GAZE-CONTINGENT COMPUTER GRAPHICS

Von der Carl-Friedrich-Gauß Fakultät
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines

Doktoringenieurs (Dr.-Ing.)

genehmigte Dissertation

von
Michael Stengel
geboren am 16. November 1985
in Magdeburg

Eingereicht am:	30. September 2016
Disputation am:	14. November 2016
1. Referent:	Prof. Dr.-Ing. Marcus Magnor
2. Referent:	Prof. Dr. rer. nat. Bernd Fröhlich

(2016)

To my loving parents.

Abstract

Contemporary digital displays feature multi-million pixels at ever-increasing refresh rates. Reality, on the other hand, provides us with a view of the world that is continuous in space and time. The discrepancy between viewing the physical world and its sampled depiction on digital displays gives rise to perceptual quality degradations. By measuring or estimating where we look, gaze-contingent algorithms aim at exploiting the way we visually perceive to remedy visible artifacts. This dissertation presents a variety of novel gaze-contingent algorithms and respective perceptual studies. Chapter 4 and 5 present methods to boost perceived visual quality of conventional video footage when viewed on commodity monitors or projectors. In Chapter 6 a novel head-mounted display with real-time gaze tracking is described. The device enables a large variety of applications in the context of Virtual Reality and Augmented Reality. Using the gaze-tracking VR headset, a novel gaze-contingent render method is described in Chapter 7. The gaze-aware approach greatly reduces computational efforts for shading virtual worlds. The described methods and studies show that gaze-contingent algorithms are able to improve the quality of displayed images and videos or reduce the computational effort for image generation, while display quality perceived by the user does not change.

Kurzfassung

Moderne digitale Bildschirme ermöglichen immer höhere Auflösungen bei ebenfalls steigenden Bildwiederholraten. Die Realität hingegen ist in Raum und Zeit kontinuierlich. Diese Grundverschiedenheit führt beim Betrachter zu perzeptuellen Unterschieden. Die Verfolgung der Aug-Blickrichtung ermöglicht blickpunktabhängige Darstellungsmethoden, die sichtbare Artefakte verhindern können. Diese Dissertation trägt zu vier Bereichen blickpunktabhängiger und wahrnehmungstreuer Darstellungsmethoden bei. Die Verfahren in Kapitel 4 und 5 haben zum Ziel, die wahrgenommene visuelle Qualität von Videos für den Betrachter zu erhöhen, wobei die Videos auf gewöhnlicher Ausgabehardware wie z.B. einem Fernseher oder Projektor dargestellt werden. Kapitel 6 beschreibt die Entwicklung eines neuartigen Head-mounted Displays mit Unterstützung zur Erfassung der Blickrichtung in Echtzeit. Die Kombination der Funktionen ermöglicht eine Reihe interessanter Anwendungen in Bezug auf Virtuelle Realität (VR) und Erweiterte Realität (AR). Das vierte und abschließende Verfahren in Kapitel 7 dieser Dissertation beschreibt einen neuen Algorithmus, der das entwickelte Eye-Tracking Head-mounted Display zum blickpunktabhängigen Rendern nutzt. Die Qualität des Shadings wird hierbei auf Basis eines Wahrnehmungsmodells für jeden Bildpixel in Echtzeit analysiert und angepasst. Das Verfahren hat das Potenzial den Berechnungsaufwand für das Shading einer virtuellen Szene auf ein Bruchteil zu reduzieren. Die in dieser Dissertation beschriebenen Verfahren und Untersuchungen zeigen, dass blickpunktabhängige Algorithmen die Darstellungsqualität von Bildern und Videos wirksam verbessern können, beziehungsweise sich bei gleichbleibender Bildqualität der Berechnungsaufwand des bildgebenden Verfahrens erheblich verringern lässt.

Summary

The discrepancy between viewing the physical world and its sampled depiction on digital displays gives rise to perceptual quality degradation. By measuring or estimating where we look, gaze-contingent algorithms aim to exploit the way we visually perceive digital images and videos to remedy visible artifacts. This dissertation presents novel gaze-contingent algorithms to enhance the perceived visual quality of conventional video footage and to improve performance when rendering virtual worlds.

This thesis starts out by highlighting fundamental background in the areas visual perception, gaze estimation, and recent research results on gaze-aware display algorithms in computer graphics.

As a first contribution, a novel gaze-aware resolution enhancement approach is presented. The algorithm allows boosting the perceived video quality beyond the actual, physical resolution of the display. The algorithm generalizes previous apparent display resolution enhancement techniques to conventional videos of arbitrary content. An optimization-based approach continuously translates each video frame in such a way that the added motion enables support for apparent resolution enhancement for the salient image region. The salient region – being congruent with the foveal viewing area allowing highest perceivable detail – is derived from an eye tracking study. The optimization algorithm takes optimal velocity, smoothness and similarity into account to compute an appropriate trajectory. For interactive guidance of the algorithm an intuitive user interface is provided. The algorithm is evaluated in a perceptual study with respect to apparent rendering quality and versatility of the method on a variety of general test scenes.

Next, the computation of perceptual motion blur in videos is presented. The technique takes the predicted eye motion into account when watching videos. Compared to traditional motion blur recorded by a video camera this approach results in perceptual blur that is closer to reality. This post-process can also be used to simulate different shutter effects for artistic purposes, or for subtle gaze direction. The proposed method handles real and artificial video input, is easy to compute and has little overhead for rendered content. A perceptual study illustrates its advantages.

The third contribution addresses the lack of reliable eye tracking in Virtual Reality (VR) headsets and precise gaze calibration. This thesis contributes an affordable hardware and software solution for drift-free eye-tracking and user-friendly lens calibration within an HMD. The use of dichroic mirrors leads to a lean design that provides full field of view while using commodity cameras for eye tracking. The prototype supports personalizable lens positioning for different inter-ocular distances. On the software side, a model-based calibration procedure adjusts the eye tracking system and gaze

estimation to varying lens positions. Challenges such as partial occlusions due to the lens holders and eye lids are handled by a novel robust monocular pupil-tracking approach. As a demonstration, a variety of gaze-aware applications are presented: gaze map estimation, accommodation simulation, gaze-contingent level-of-detail, gaze control of virtual avatars, and gaze analysis for immersive videos.

With increasing display resolution for wide-field-of-view VR headsets, shading has become the major computational cost in real-time rendering. Therefore, the fourth and last contribution addresses gaze-contingent rendering. To reduce computational effort, an algorithm is presented that only shades visible features of the image. The remaining image pixels are cost-effectively interpolated without affecting perceived quality. In contrast to previous approaches the novel perceptual method introduces a flexible sampling scheme that incorporates multiple aspects of the human visual system: acuity, eye motion, contrast (stemming from geometry, material or lighting properties), and brightness adaptation. The sampling scheme is incorporated into a deferred shading pipeline to shade perceptually relevant fragments of the image while a pull-push algorithm interpolates the radiance for the remaining pixels. The approach does not impose any restrictions on the performed shading. Conducted psycho-visual experiments validate scene- and task-independence of the method. The number of fragments that need to be shaded is reduced by 50 % to 80 %. Importantly, the algorithm scales favorably with increasing resolution and field of view, rendering it well-suited for VR headsets and wide field of view projection.

The dissertation concludes with a discussion on future directions of gaze-aware displays and extensions of the presented approaches.

Zusammenfassung

Die Differenz zwischen der physischen Welt und deren digitale Repräsentation führt zu wahrnehmbaren Qualitätseinbußen bei der Darstellung auf dem Bildschirm. Die Bestimmung der Blickrichtung (Eye Tracking) erlaubt es blickpunktabhängigen Algorithmen, unter Ausnutzung der Grenzen der menschlichen Wahrnehmunug sichtbare Artefakte zu reduzieren. Im Rahmen dieser Dissertation werden neuartige Algorithmen für blickpunktabhängige Displays vorgestellt, um die wahrgenommene visuelle Videoqualität zu erhöhen und das Rendern virtueller Welten zu beschleunigen.

Einleitend werden die Grundlagen der visuellen Wahrnehmung und Blickrichtungserfassung sowie darauf bezogene forschungsrelevante Arbeiten im Bereich Computergraphik dargestellt.

Als erster Beitrag der Arbeit wird ein Algorithmus zur wahrnehmungsbasierten Auflösungserhöhung vorgestellt. Dieser erlaubt es, den wahrgenommenen Detailgrad von Videos über die physische Displayauflösung hinaus zu erhöhen. Der Algorithmus stellt eine Verallgemeinerung vorheriger software-basierter Methoden zur wahrnehmungsbasierten Auflösungserhöhung für beliebiges Videomaterial dar. Der Ansatz baut auf einem Optimierungsverfahren auf, mithilfe dessen eine kontinuierliche Translation eines jeden Videobildes herbeigeführt wird. Diese Bewegung resultiert in einer Maximierung des Auflösungserhöhungseffektes in jenem Bereich des Videobildes, der die höchste Salienz aufweist und sich daher mit größter Wahrscheinlichkeit im fovealen Sichtbereich befindet. Die Berechnung der Salienz des Videos erfolgt auf Basis zuvor gemessener Blickdaten. Der Optimierungsalgorithmus nutzt Informationen über die optimale Bewegungsgeschwindigkeit, die Glattheit der generierten Bewegung und die Nähe zur Ausgangsposition des Videobildes, um eine geeignete Bewegungstrajektorie zu ermitteln. Zur interaktiven Unterstützung des Algorithmus wurde eine geeignete Benutzeroberfläche umgesetzt. Der Algorithmus wurde im Rahmen einer Wahrnehmungsstudie evaluiert. Dabei wurden Erkenntnisse über die resultierende Videoqualität und die Anwendbarkeit des Verfahrens für allgemeine Videoinhalte gewonnen.

Im zweiten Beitrag der Dissertation wird ein Berechnungsmodell zur Generierung der perzeptuellen Bewegungsunschärfe in Videos vorgestellt. Die neue Technik filtert auf Basis eines Wahrnehmungsmodells zeitlich entlang des geschätzten Blickpfades, einer Abfolge aus Fixationspunkten, den ein Zuschauer bei der Betrachtung eines Videos im Mittel vollführt. Im Vergleich zu herkömmlichen Videos, in denen die Bewegungsunschärfe durch die Kamera selbst aufgezeichnet wurde, erlaubt der perzeptuelle Filter eine natürlichere und realitätsnähere Darstellung bei der Betrachtung des Videos. Der Filtermechanismus kann zusätzlich zur Vermeidung von wahrnehmbaren Videoartefakten wie zum Beispiel Ghosting im periphären Sichtbereich auch zur Simulation unterschiedlicher Shutter-

verfahren eingesetzt werden oder sogar den Blick des Betrachters auf subtile Weise zu lenken. Die Methode erlaubt das Filtern sowohl von realen wie auch von künstlich generierten Videos und ist zudem effizient berechenbar. Das Verfahren wird im Rahmen einer Wahrnehmungsstudie evaluiert.

Der dritte Beitrag resultiert aus der Beobachtung, dass mobile VR displays (Head-mounted Displays, HMDs) in Verbindung mit Eye Tracking wenig verbreitet sind. Zusätzlich ist die Kalibrierung dieser Geräte schwierig. Zu diesem Zweck wird in diesem Teil der Arbeit eine kostengünstige Hardware- und Softwarelösung zur driftfreien Blickrichtungserfassung und zur praktikablen Kalibrierung des HMDs vorgstellt. Die Integration dichroitischer Spiegel ermöglicht den Erhalt des vollen Blickfeldes des Nutzers und zudem den Einsatz kostengünstiger Kameras normaler Größe für das Tracking. Der erstellte Prototyp erlaubt die Anpassung der Linsen an die Sehfähigkeiten des Trägers des HMDs. Softwareseitig ermöglicht die modellbasierte Kalibrierungsprozedur die Blickrichtungserfassung für variable Linseneinstellungen. Partielle Verdeckungen durch die Linsenhalterungen und die Augenlider des Benutzers werden bei der robusten monokularen Pupillenerfassung berücksichtigt. Die Vorteile und die Funktionalität des HMD-Prototypen werden für verschiedene Anwendungen demonstriert. Dazu gehören die Erstellung von Blickpunktkarten (Gaze map), Akkommodationssimulation des Auges, blickpunktabhängige Renderqualität, Blickanimation virtueller Avatare und die Blickanalyse in immersiven Videos.

Aufgrund der steigenden Auflösung von HMDs stellt die Berechnung des Shadings im Bereich des Echtzeitrenderns den rechenintensivsten Schritt dar. Der vierte und letzte Beitrag untersucht daher das blickpunktabhängige Rendern für Echtzeitanwendungen (Gaze-contingent Rendering). Das Resultat ist ein Algorithmus, der den Berechnungsaufwand beim Render reduziert, indem ausschließlich die bei der Wahrnehmung ausschlaggebenden visuellen Merkmale ausgewertet werden. Die übrigen Bildpixel werden ohne größeren Berechnungsaufwand interpoliert. Im Vergleich zu bisherigen Rendertechniken erlaubt das neue Verfahren ein flexibles Sampling entsprechend eines effizient auswertbaren Wahrnehmungsmodells. Komponenten des Modells umfassen die räumlich-zeitliche Wahrnehmungsschwelle, Kontrast auf Basis von Szenengeometrie, Material und der Beleuchtungssituation, sowie die zeitbedingte Helligkeitsadaption. Das Samplingschema ist in ein modernes Deferred Shading Renderverfahren eingebettet. Ein effizienter Pull-Push Schritt ermöglicht nach dem Shading der perzeptuell ausgewählten Bildpixel die Komplettierung der noch fehlenden Bildteile. Der Ansatz erhebt dabei keine Ansprüche an das verwendete Shadingverfahren. Die Unabhängigkeit von der zugrundeliegenden Szene und von kognitiven Faktoren wurden im Rahmen einer psychophysiologischen Studie für eine Auswahl von Testszenen bestätigt.

Die Anzahl der vollständig darzustellenden Bildpixel wird durch das Verfahren um 50 bis 80 % reduziert. Eine wesentliche Eigenschaft des Verfahrens ist die sublineare Skalierung entsprechend der Bildauflösung und des Blickbereiches, wodurch die Methode insbesondere für zukünftige HMDs mit großem Blickbereich und hoher Auflösung einen steigenden Laufzeitvorteil bedeutet.

Die Dissertation schließt mit der Diskussion weiterführender Arbeiten im Bereich blickpunktabhängiger Bildschirme und der Erweiterungsmöglichkeiten der dargebrachten Methoden.

Acknowledgements

Many people supported and inspired me during the work on my thesis. First and foremost I am grateful to my supervisor Prof. Marcus Magnor. I enjoyed working at the TU Braunschweig together with you. You have shown me interesting new research directions, gave me the freedom to pursue my own ideas and motivated me for the major conference deadlines. I am also deeply grateful for the many conferences I was able to visit during that time.

I would especially like to thank all my colleagues that have worked with me on the proposed publications, in particular Martin Eisemann, Steve Grogorick, Christian Lipski, Lorenz Rogge, Felix Klose, Pablo Bauszat, Benjamin Hell, Kai Ruhl, Kai Berger, Thomas Neumann, Jan-Philipp (JP) Tauscher and Thomas Löwe. It has been both very fruitful and a great pleasure working with these splendid researchers. Thanks to all members of the Computer Graphics Lab—Thiemo, Matthias, Emmy, Georgia, Anja, Skrollan and Carsten—for the discussions, help and for making it such a great environment to work at. Special thanks to JP for proof-reading the draft of this dissertation. I also like to thank all the people who participated in the various studies and video recordings for the projects. Additional thanks go to Ariana Prekazi and Franziska Ludwig for supporting me in conducting perceptual studies.

I would like to thank Krzysztof Templin, Stephen Higgins, Evin Grant, Keefa Chan, Alexandre Pestana, Frank Meinl, Morgan McGuire and Andrew Maximov for granting permission to use their code and data in my studies. Thanks to the many companies, Epic Games, Inc., Crytek GmbH, NVIDIA Corp. and Blender Foundation, for their openly available game engines and stock footage. Thanks also goes to The Foundry Ltd. for providing Nuke software licenses for academic use. Special thanks is due to Steve Grogorick who worked with me relentlessly on the head-mounted display prototype and the source code to make the deadlines. Big thanks to my students Marc Kastner, Andreas Bauerfeld, Inga Menke and Sascha Fricke who have been great persons to discuss and try out ideas or just to spend some quality time during coffee breaks.

I am most grateful to my parents Günther and Birgit and to my brothers Robin and Andreas as well as to Miriam. You have always supported me and gave balance to my academic life. Ela, thank you for your encouragement, support and motivation—thanks for always being there for me!

The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 256941, *Reality CG*, and the DFG Reinhart Kosselleck Project *Immersive Digital Reality*.

Contents

Preface

This dissertation is the result of several publications I have authored in cooperation with different collaborators. My advisor Marcus Magnor is co-author on all of my publications. In this thesis the publications are presented in the common context of gaze-contingent display methods.

In the following I clarify my individual contributions to the respective publications. The publications are ordered according to the structure of my dissertation.

- Michael Stengel, Martin Eisemann, Stephan Wenger, Benjamin Hell, and Marcus Magnor.
  Optimizing Apparent Display Resolution Enhancement for Arbitrary Videos.
  In *IEEE Transactions on Image Processing (TIP)*, vol. 22, no. 9, pages 3604–3613, September 2013. Part of the project *Reality CG*. Patent number 10 2013 105 638.
  My part on this project was to implement the trajectory optimization algorithm and the high frame rate video player. In addition, I developed the interactive video enhancement tool and conducted the eye tracking and perceptual study. S. Wenger suggested usage of an Expectation-Maximization algorithm for the trajectory optimization. B. Hell contributed a derivation for faster convergence of the optimization problem for trajectory computation. M. Eisemann accompanied the project with many helpful suggestions. M. Eisemann and I worked closely together while writing the paper. M. Magnor gave advice concerning the results and content of the paper.

- Michael Stengel, Pablo Bauszat, Martin Eisemann, Elmar Eisemann, and Marcus Magnor.
  Temporal Video Filtering and Exposure Control for Perceptual Motion Blur.
  In *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, vol. 21, no. 5, pages 663–671, May 2015. Part of the project *Reality CG*.
  I have contributed the idea of including saliency-based gaze prediction for a gaze-contingent temporal video filter. I also wrote most parts of the implementation of the temporal filter. In addition, I have generated and processed all of the test scenes used in the study. I also conducted and analyzed the perceptual experiments on subtle gaze direction. P. Bauszat was very supportive in discussing ideas for the perceptual filter and helped in improving the implementation to achieve interactive filtering rates. M. Eisemann initiated the idea of temporally filtering videos and was helpful in discussing correctness of the filter. M. Eisemann, E. Eisemann and I worked closely together on the conference paper. M. Magnor gave advice concerning the content of the paper. E. Eisemann contributed the synthetic rolling shutter results for the paper.

- Michael Stengel, Steve Grogorick, Martin Eisemann, Elmar Eisemann, and Marcus Magnor.
  An Affordable Solution for Binocular Eye Tracking and Calibration in Head-mounted Displays. In *Proc. of ACM Multimedia*, pages 15–24, October 2015. Part of projects *Reality CG*.
  The paper was selected as the best student paper on the ACM Multimedia '15 conference. For this project I worked with S. Grogorick. Being his supervisor, Steve worked on different aspects of the eye tracking head-mounted display (ETHMD) for his master thesis, such as the pupil detection algorithm, the 3d-printable prototype, and an ETHMD plugin for the Unreal Engine. Together, Steve and I created and assembled the prototype. I have contributed ideas for the nonobscuring tracking within the HMD, for the pupil tracking algorithm and contributed a framework for calibration and gaze mapping. In addition, I contributed applications and results and supervised optimization of the pupil tracking algorithm performed by my student Marc Kastner. M. Eisemann and I worked together on the conference paper. B. Hell gave support with respect to the notation of the contributed pupil tracking algorithms. M. Magnor and E. Eisemann gave advice concerning the content of the paper.

- Thomas Löwe, Michael Stengel, Emmy-Charlotte Förster, Steve Grogorick, and Marcus Magnor.
  Visualization and Analysis of Head Movement and Gaze Data for Immersive Video in Head-mounted Displays. In *Proceedings of the Workshop on Eye Tracking and Visualization (ETVIS)*, vol. 1, October 2015. Part of project *Scalable Visual Analytics*.
  My contribution on this project was the initial idea of gaze analysis in immersive video. In addition, I have implemented the video player for the ETHMD and conducted the perceptual study. Implementation of the visualization framework for gaze analysis has been mostly contributed by T. Löwe, Steve Grogorick and Emmy-Charlotte Förster. I have contributed ideas for the layout of the user interface and implemented the heat map visualization. T. Löwe, Emmy-Charlotte Förster and I worked closely together on the workshop paper. M. Magnor gave advice concerning the content of the paper.

- Michael Stengel, Steve Grogorick, Martin Eisemann, and Marcus Magnor.
  Adaptive Image-Space Sampling for Gaze-Contingent Real-time Rendering.
  In *Computer Graphics Forum (Proc. of Eurographics Symposium on Rendering EGSR)*, vol. 35, no. 4, pages 129–139, July 2016. Part of the projects *Immersive Digital Reality*, and *Reality CG*.
  The paper was selected as the best paper on the EGSR '16 conference. For this project I collaborated with S. Grogorick. Together we worked on many ideas for the perceptual sampling and on the implementation of the rendering framework. I have derived and implemented different components of the perceptual model, such as the texture adaptation, brightness adaptation, object saliency and motion dependency. I have also implemented the physically-based shading approach and debugging tools. In addition I have contributed test scenes and prepared the perceptual study. I also contributed the fast fragment discard function. S. Grogorick implemented the stereo mode of the framework, the pull-push operation, the acuity fall-off shader and the sampling generation. Together S. Grogorick and I conducted the perceptual study. I have contributed the analysis of the

results. In addition, S. Grogorick and I contributed several improvements and extensions of the ETHMD, such as a higher-resolution display and improvements on the gaze tracking framework. I have written most parts of the paper. M. Eisemann gave support in improving the conference paper and shared ideas on the perceptual model. M. Magnor gave advice concerning the paper.

- Michael Stengel and Marcus Magnor. **Gaze-contingent Computational Displays**.
  In *IEEE Signal Processing Magazine (SPM)*, vol. 33, no. 5, pages 139–148, September 2016. Part of the projects *Immersive Digital Reality*, and *Reality CG*.
  M. Magnor initiated the idea of contributing a comprehensive article on recent trends in gaze-contingent algorithms and my contributions in this field. I have contributed large parts of the final work. M. Magnor and I worked together on the article.

I have co-authored additional publications which are not included in this thesis:

- Jan Jacobs, Michael Stengel, and Bernd Fröhlich. **A Generalized God-Object Method for Plausible Finger-Based Interactions in Virtual Environments**. In *Proceedings of the IEEE Symposium on 3D User Interfaces (3DUI)*, pages 43–51, March 2012. *The paper was selected as the best paper on the 3DUI '12 conference.*

- Lorenz Rogge, Felix Klose, Michael Stengel, Martin Eisemann, and Marcus Magnor. **Garment Replacement in Monocular Video Sequences**. In *ACM Transactions on Graphics (TOG)*, vol. 34, no. 1, pages 6:1–6:10, November 2014.

- Anna Hilsmann, Michael Stengel, Lorenz Rogge. **Cloth Modeling**. Book chapter in *Digital Representations of the Real World: How to Capture, Model, and Render Visual Reality* by Marcus Magnor, Oliver Grau, Olga Sorkine-Hornung and Christian Theobalt, pages 229–243, May 2015.

Chapter 1

Introduction

Contents

1.1 Motivation

Display technology is advancing at a breath-taking pace. Driven by consumer market demands, screen size, resolution, contrast and refresh rates are growing bigger, higher and faster almost on a weekly basis. Nevertheless, the fundamental difference between the continuous physical world and its digitally sampled and displayed image still gives rise to perceptually noticeable quality degradation. If matched to eye acuity so that single pixels cannot be perceived anymore, common Full-HD screens result in a narrow vertical field of view (FOV) of 18° [SM16]. For a more immersive viewing experience high-end screens with 5k or 8k resolution are required to widen the FOV. Exceptions are smartphone displays that do feature high pixel densities of 500 to 800 pixels per inch (ppi) to enable clear readability of small-scale text and playback of high-resolution videos. This implies, however, that more and more pixels must be rendered on resource-limited mobile devices. Unfortunately, the provided video bandwidth and rendering performance increases at a lower pace than what can be displayed in terms of the total number of pixels.

Incongruities also exist in the temporal domain between digital video recording and display capabilities. While TV display refresh rates today commonly match or exceed 60 Hz, standard video acquisition frame rates still hover between 24 and 30fps. The discrepancies between the physical world and its digital representation, as well as the mismatch in acquisition vs. display capabilities and displays vs. human visual perception, can cause noticeable artifacts.

Gaze-contingent display (GCD) algorithms basically follow a simple idea. If it is known, or can be reliably estimated, how the human visual system perceives digital images at any time, gaze-contingent display methods are able to make use of a number of perceptual strategies to improve perceived visual quality beyond the limits of the physical devices. In addition, gaze contingency allows allocating computational resources on-the-fly to image regions that are perceptually relevant for the current gaze direction, resulting in reduced bandwidth or processing requirements while keeping rendering quality perceptually lossless compared to traditional full-resolution rendering. The perceptual importance of a given image or video is either described by saliency and task maps or directly estimated by tracking the gaze of the user.

The notion of GCD devices dates back at least two decades. GCD research touches various fields, including computer graphics, computer vision, visualization, psychology as well as neuroscience. Respective computational algorithms enable a great variety of applications in research, industry and entertainment. A taxonomy of state-of-the-art gaze-contingent algorithms based on properties of the plenoptic function [1] such as contrast, color gamut, spatial resolution, temporal resolution, and angular resolution, is given in Masia et al. [MWDG13].

While gaze-contingent display approaches have been proposed before, only recently have eye tracking hardware, saliency estimation methods and graphics hardware become sufficiently fast,

[1] The plenoptic illumination function by Adelson et al. is an idealized model to express a 2D image of a scene from any possible viewing position at any viewing angle at any point in time [AB91].

robust and affordable to allow for incorporating advanced gaze-aware methods in mass-market devices [New16]. However, only few gaze-aware methods have so far been adopted to VR headsets due to lack of available gaze-tracking hardware.

This dissertation proposes novel gaze-contingent computational display approaches suited for video playback and real-time VR rendering, enhancing perceived visual fidelity of common, consumer-market display technologies. The proposed techniques have been inspired by the success of previous gaze-contingent displays that show great potential to enhance the visual quality of computer graphics by fruitful combinations of perceptual considerations, computational display algorithms and minor hardware modifications [OHM+04, DÇ07, MWDG13].

1.2 Topics and Contribution

The ideas described in this dissertation have been published in international journals and conference proceedings. An overview of the proposed contributions will be published in the Special Issue on Computational Photography and Displays of the IEEE Signal Processing Magazine [SM16]. The main parts and corresponding contributions of the dissertation are summarized in the following.

- **Boosting Spatial Resolution**

  The required down-sampling from spatially high-resolution video footage to the resolution of a display device results in a loss of fine details and an overall reduction of perceived image quality. Modern smart phone displays and selected desktop monitors, such as Apple's Retina Display™, achieve eye resolution limit. However, this is not true for most display devices, particularly for VR headsets. Screens for VR headsets must be small in size but at the same time cover a very wide FOV. Even with state-of-the-art 350 ppi screens, current HMD displays are still an order of magnitude away from eye acuity, exacerbated even more by significant pixel magnification in the central region of lens-based HMDs.

  Super-resolution displays mostly require specialized hardware configurations, such as regular vibrations of the screen [BF12b]. Alternatively, apparent display resolution enhancement techniques have been recently proposed to provide purely software-based super-resolution on high refresh-rate displays [DER+10a, TDR+11]. By exploiting how the Human Visual System observes and processes moving content, these approaches are able to boost perceived resolution beyond the actual, physical resolution of the display. With active-matrix organic light-emitting diode technology, refresh rates in excess of 1000 Hz are achievable, allowing for a sixfold increase in apparent display resolution. VR headsets may therefore significantly benefit from super-resolution techniques. In essence, apparent display resolution enhancement allows trading screen refresh rate for perceived resolution as long as the user's gaze continuously and predictably tracks the moving foreground via smooth pursuit eye movement.

  This thesis introduces an approach that allows for apparent resolution enhancement even for scenes that originally do not contain any movement or for which optical flow computation is difficult

or impossible. Based on the assumption that our gaze follows the most salient regions of the sequence, in a pre-process the salient foreground regions are determined. The video frames are then continuously and unnoticeably shifted in such a way that, in combination with the postulated smooth pursuit eye movement, the tracked foreground moves in a diagonal direction, facilitating the exploitation of the resolution enhancement effect, Chapter 4.

- **Boosting Temporal Fidelity**

  In the physical world our individual gaze determines if, how, and where we perceive blur. Our blur perception in the real world can differ distinctly from camera-recorded motion blur. While watching live-action shots on screen, we may notice annoying ghosting, judder, or edge banding artifacts [Abr13, Fen14].

  These judder artifacts become especially apparent when viewed on wide-angle displays because spurious high-frequency details in the moving background lead to the perception of discontinuous, jaggy motion by our peripheral vision. Because our peripheral vision is especially sensitive to movements and expects consistent, smooth motion, the discontinuous motion of the background distracts our visual attention from the tracked foreground.

  A straightforward solution to reduce jaggy motion is to use higher frame rates. Indeed, new displays work with higher refresh rates. However, video content is still broadcast at a low frame rate (24-30 fps) in the consumer market environment due to limited bandwidth of used media. Therefore, TV manufacturers try to solve the problem by frame interpolation [Fen06]. However, required optical flow computations are prone to errors for fast motion [DER+10b].

  To overcome such limitations, in this thesis a software-based technique is proposed which replicates the temporal summation behavior of the human visual system (HVS) based on eye-tracking and saliency data created for the movie. Based on interpolated ultra high-frame rate videos, the technique filters the movie by introducing the correct amount of blur on a per-pixel basis that is required for a given output frame rate. When watching the filtered video, hold-type blur and other perceivable artifacts are greatly reduced, Chapter 5.

- **Gaze Tracking in VR Headsets**

  A renaissance of Virtual Reality (VR) and Augmented Reality (AR) can be observed due to the success of high-quality VR/AR headsets such as the Rift™ from Oculus VR, HTC Vive™ or Microsoft Hololens™ and also low-cost solutions such as the Google Cardboard™. However, the display quality in terms of spatial resolution and dynamic range is still far from optimal. Although computational power of GPUs grows quickly, the development cannot compensate for the demands of increasing screen resolution, refresh rate and scene complexity required for high-fidelity rendering. As an alternative, perceptually-inspired rendering methods offer an attractive alternative. VR developers also face other challenging problems. One recurrent issue is the need for precise calibration to correctly adjust disparity and parallax [JSIS+08]. Missing calibration and the well-known vergence-accommodation conflict (VAC), which happens when presenting

stereoscopic 3D content with a fixed-focus distance of the screen, result in fatigue and other forms of visual discomfort [HGAB08]. A decrease in immersion can also be observed in indirect interaction concepts and restricted inter-subject communication solutions due to limitations in tracking and corresponding constraints in avatar animation.

Real-time gaze tracking in the VR headset can be helpful for all of these problems. Calibration is much easier and offers higher precision in case the locations of the eyes can be measured. Real-time eye tracking also allows reducing VAC by adjusting vergence and simulating accommodation on the fly. Avatars can be rendered much more naturally if gaze is animated which increases the possibility to overcome the *Uncanny Valley* and enhances immersion by enabling collaborating users to establish eye contact in VR. Last but not least, and maybe most importantly, the amount of rendered pixels can be adjusted to the resolution of the HVS. This is known as *foveated rendering* and results in a fraction of the rendering cost. The notion of foveated rendering is also useful for broadcasting and for rendering of immersive 360° videos. Currently, video codecs are optimized for encoding blocks of pixels at the same resolution in every part of the video frame. In light of the retina's vastly varying perception characteristics from foveal to peripheral vision, however, future gaze-contingent video codecs may be able to adapt coding rate to local view eccentricity. With gaze-contingent video encoding, only perceptually relevant information needs to be transmitted and rendered, saving bandwidth and memory.

Although expensive solutions for VR headsets exist, gaze-tracking has not yet been widely investigated due to the cost and hardware complexity. This thesis contributes a novel gaze-tracking head-mounted display based on a mirror-based setup.In addition, a precise low-latency tracking algorithm and a simple and efficient calibration method for the proposed hardware setup are presented. The overall objective of the proposed gaze-aware VR headset is to exploit properties of the HVS for immersive displays, leaving limitations of current HMDs behind for a completely immersive experience, Chapter 6.

- **Boosting Rendering Performance**

  With knowledge about where visual attention will be directed in an image or video, only perceptually important regions need to be rendered with high quality. Therefore, rendering time does not have to be spent on perceptually less important regions. However, it is crucial for a gaze-contingent rendering method to achieve image quality that is perceptually indistinguishable from a fully converged solution.

  This thesis contributes a gaze-contingent render pipeline for accelerated real-time rendering in VR environments with a wide FOV. Incorporating visual cues such as acuity, eye motion, adaptation and contrast into a single flexible perceptual model, the algorithm employs a perceptually-adaptive sampling pattern which is used for sparse shading of the scene to be rendered. Efficient image interpolation creates an image of the same perceived quality as if shading each fragment, but at a fraction of the original shading costs. The resulting image contains high object detail in the foveal

region and increasingly less detail towards the periphery. It has been proven to be indistinguishable from a fully shaded reference. Therefore, the proposed algorithm is suitable for optimizing the visual experience of full FOV immersive displays by simultaneously accommodating the characteristics of the HVS across the full visual field, Chapter 7;.

1.3 Dissertation Organization

Chapter 2 conveys an understanding of perception mechanisms of human vision. The following Chapter 3 describes recent work on gaze-contingent display algorithms and applications being related to contributions proposed in this thesis. Due to the large body of literature, however, only the most relevant publications of existing literature can be covered. Next, two novel display methods are presented to enhance perceived visual quality of conventional video footage when viewed on commodity monitors or projectors. Chapter 4 covers spatial resolution enhancement, whereas temporal fidelity is covered in Chapter 5. For gaze-aware Virtual Reality, a novel head-mounted display with real-time gaze tracking is described in Chapter 6, as well as different applications in the context of Virtual Reality and Augmented Reality. A novel gaze-contingent render method is described in Chapter 7 that greatly reduces computational effort for shading virtual worlds. The dissertation concludes with an appraisal of the contributed techniques in Chapter 8.

Chapter 2

Background

Contents

This chapter presents the background on the human visual system (HVS) and visual perception being important for the methods covered in this dissertation. Relevant sections of this background chapter are referred to in the remaining parts of the thesis.

Although correlations between physical stimuli and perceived information are not yet fully understood, models of the HVS enable us to conservatively express different important features of human vision. Due to the vast amount of research during the last centuries, the section will cover only those aspects of visual perception which are of concern for the gaze-contingent approaches proposed in this dissertation.

2.1 Introduction to Human Visual Perception

The goal of this section is to provide an understanding of human vision with respect to perceptually-motivated computer graphics methods. The chapter focuses on the limits of visual performance, especially on visual acuity. Although human vision is a highly non-linear system that is not yet understood completely, selected aspects can be modeled successfully. Interestingly, some of those models can be described by linear functions rendering them ideally suited for efficient implementation in real-time computer graphics applications.

The first part of the section will briefly describe low-level features of the *visual system* beginning with *eye physiology* and the capturing of light in this complex organ. Features of different cells in the eye and in the brain are described and their importance for low-level vision.

In the next part, concepts of *visual acuity* are outlined. The quality of vision is highly dynamic across the visual field and across the spectrum of environmental conditions. Modeling visual acuity forms the starting point to exploit limitations in perceivable image detail.

Humans sense the intensity of light, but they are much more sensitive to contrast. For example, when the sun is occluded by clouds, we see the changing intensity of light, but much more do we perceive the difference of brightness between one object and another. In order to benefit from this and related observations, the perception of *spatial contrast* is discussed in greater detail.

Interestingly, we are able to perceive contrast over a huge range of intensities. Why this works and which parts of the HVS are involved is described in the section on *adaptation*.

Our brain integrates discrete snapshots of objects and actors into dynamic perceptual events. Discussions on *temporal contrast* and *motion processing* explain dependencies, requirements and limits in human perception of a dynamic world.

Gaze contingency can only be achieved if gaze direction can be derived. The orientation of both eyes is a strong hint for gaze. However, the gaze direction does not change arbitrarily but is constrained by physical limitations and influenced by higher-level vision. Resulting effects and ways for taking those into account computationally are covered in the parts on *eye motion* and *attention*.

A list of facts describing the most relevant properties of visual perception for the area of gaze-contingent algorithms summarizes the section.

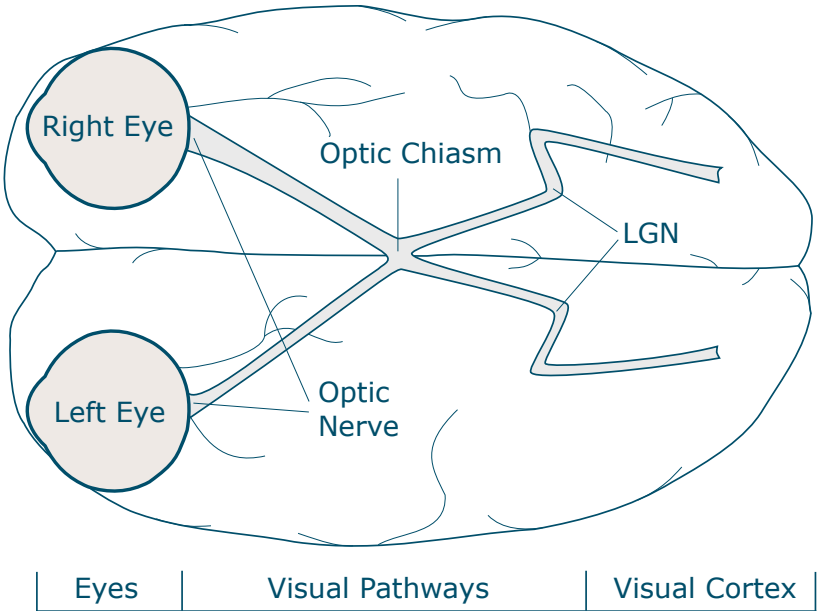

Fig. 2.1 Major parts of the Human Visual System. *Image after* [Lue03]

2.2 The Visual System

Our visual system contains at least three processing parts: the *eyes*, the *visual pathways* and the *visual cortex*. The eyes receive incoming light reflected from the environment. They convert collections of photons into electrical signals which are then transported along the visual pathways [Lue03]. The visual cortex of the brain interprets the incoming signals and enables visual perception. Since the eye and the visual cortex contribute most to the vision process they are discussed in detail in the following sections.

Eye Physiology

The human eye is a complex optical instrument which has been tuned precisely to life on earth by evolution. The spherical eye ball consists of a focusable lens, an adjustable aperture (*iris* and *pupil*), a photodetector layer (the *retina*) and fluids supporting its optical requirements (*aqueous humor* and *vitreous humor*). In the following the most important components of the eye are briefly explained. First, *visible light* [1] travels from the outermost layer to the innermost surface of the eye, called the retina. With a thickness of 0.55 - 1.0 mm and about 12 mm in diameter, the sclera forms a protecting layer for the eye. Toward the front of the eye the mostly white tissue turns into the transparent *cornea* which allows photons to enter the eye. The cornea has a index of refraction (IOR) close to that of

[1] The human eye is tuned to respond to light ranging from 370 to 730 nm in wavelength, consequently referred to as *visible light*.

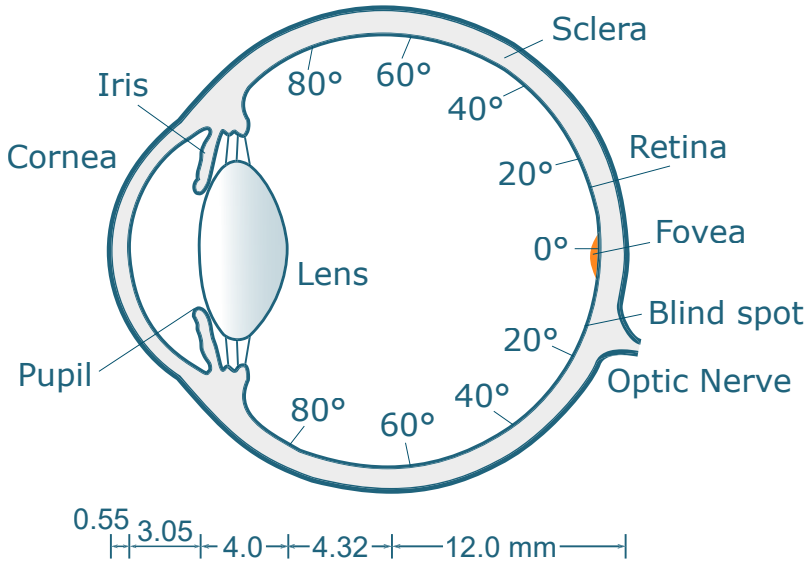

Fig. 2.2 Average dimensions of the eye. The angular eccentricity values indicate locations in degrees visual angle relative to the fovea. *Image after* [AKLA11]

water and has a convex surface resulting in the most powerful focusing element of the eye.[2] From the cornea the light enters the *anterior chamber* which is filled with a water-like substance, the *aqueous humor*. This chamber is confined by the *iris*, a muscle with a central aperture known as the *pupil*. Pigments distributed in the iris result in characteristic eye colors. With contraction of the iris the size of the pupil changes and, correspondingly, the amount of light entering the eye. From a maximum diameter of 8 mm in lowest light levels the pupil diameter may decrease to less than 2 mm at sunlight. The pupil allows access of the light to a crystalline *lens* with dynamic optical power. The lens contains a unique protein concentration which results in material of high refractive index and transparency. The action of the surrounding *ciliary muscles* permits the lens to increase or decrease power and therefore allows the eye to focus at different distances. The *accommodation* mechanism is explained in greater detail in Sec. 2.4. After the light is refracted by the lens, it enters the inner part of the eye, called the *posterior chamber*. This chamber is filled with a transparent fluid, the vitreous humor, and contains two layers of tissue: the *choroid* and the *retina*. Caused by its pigmentation, the choroid forms a layer that reduces light scattering inside the posterior chamber. After passing through the cornea, the aqueous humor, the lens and the vitreous humor, light is focused onto the innermost layer, the retina. This layer of transparent tissue covers the back of the eyeball and consists of photoreceptors and neurons to collect photons and to process the resulting impulses. Many aspects of vision result from the physiology of the retina. Therefore, this layer is explained in greater detail in the following.

[2] Refractive indices: air 1.000; glass 1.520; water 1.333; cornea 1.376. Optical power (diopters): cornea 43; lens (relaxed) 20; whole eye 60. [Wan95]

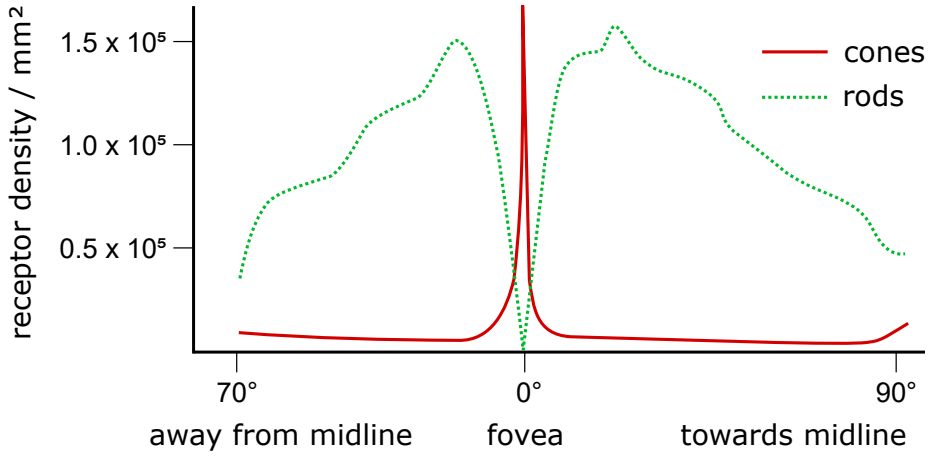

Fig. 2.3 Photoreceptor distribution. The curves show the eccentricity-dependent density distribution of rods and cones in the human retina. *Image after* [Ost35]

The Retina

The retina contains photoreceptor cells that convert incoming light energy into neural signals. The neural signals are gathered by connected *collector cells* and filtered by the higher-level *retinal ganglion cells*. In the *optic nerve* the filtered signals leave the retina and are transported to the visual cortex via the visual pathways.

Retinal photoreceptor cells come in two principal classes: *rods* and *cones*.

- Rods absorb light over a broad spectral range with high sensitivity. They provide monochromatic vision under low light-level conditions (*scotopic vision*), e.g. at night. At daylight, the rods are permanently saturated and therefore deactivated.
- Cones are responsible for sharp color vision under normal daylight conditions (*photopic vision*). The denser the cone receptors are packed, the more acute the vision.

Rods and cones contain the molecule *rhodopsin* which absorbs visible light and eventually triggers an electrical nerve signal. The signal is then transported to the higher-level ganglion cells. This event occurs in less than one millisecond. In total, the human retina contains about 90 to 200 million rods and about 4 to 8 million cones. However, the optic nerve contains only about 1 million individual fibers and an equal amount of connected retinal ganglion cells [Fai13]. The raw visual information received by the photoreceptors must therefore be downscaled to a more compact, non-uniform *retinal image* by the ganglion cells before it is transported further via the visual pathways.

For photopic vision, three distinct classes of cones exist: *S-cones*, *M-cones* and *L-cones*. The distribution of S-, M- and L-cones follows a ratio of roughly 1:5:10 [TFCRS11, p.130]. Each class of cones is sensitive to a limited spectral range and has its peak sensitivity at a different wavelength.[3]

[3] Three types of cones let humans be classified as *trichromats*.

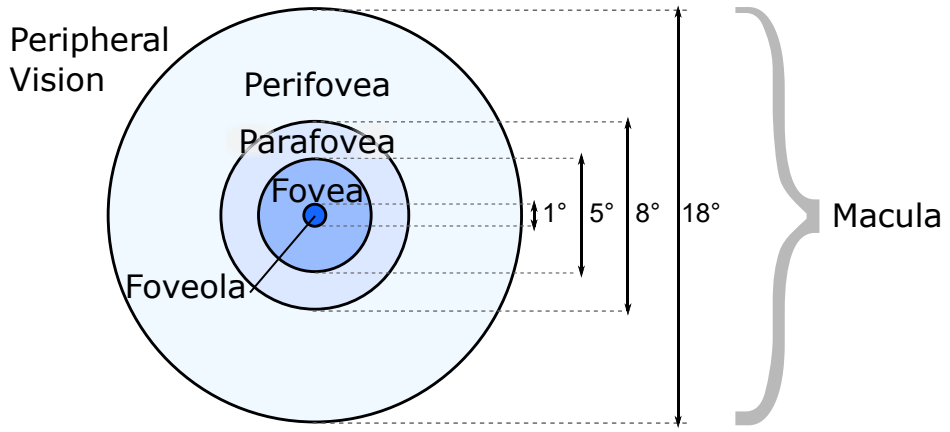

Fig. 2.4 Foveal zones with eccentricity values. *Classification according to* [Wan95].

S-cones have a peak sensitivity at 440 nm whereas M-cones and L-cones reach peak sensitivity at 530 nm and 560 nm, respectively. The combined responses of all the cone types allows us to distinguish tens of thousands of different color hues. On the retina still, output of the cones is re-encoded into tree opponent values: red-green, blue-yellow and a brightness value [DVDV93].

Photoreceptors are not uniformly distributed across the retina. The packing density of both rods and cones varies dramatically over the retina (see Fig. 2.3). The cones reach their highest density at the fovea with a peak density of approx. 1.6×10^5/mm². In this area cones are distributed in an almost regular, dense arrangement. The higher the packing of cones the higher the spatial resolution of the retinal image. S-cones are completely absent from the fovea. Hence, fine spatial details are received only by using M- and L-cones. Rods are completely missing from the fovea. The transition between the fovea and its periphery is smooth and there is no well-defined boundary in between [SRJ11]. Commonly, the *foveola* (also fovea centralis) covers 1° of visual angle [Wan95]. The *fovea* extends to 5° (Fig. 2.4). The *parafovea* (5 - 8°) and the *perifovea* (8 - 18°) extend around the fovea. Together the foveal regions make up the *macula*. The peripheral vision follows from 18° up to 180° horizontal visual angle. For this dissertation *foveal vision* refers to eccentricities $< 2°$, the *central visual field* to $< 8°$ and *peripheral vision* to larger eccentricities.

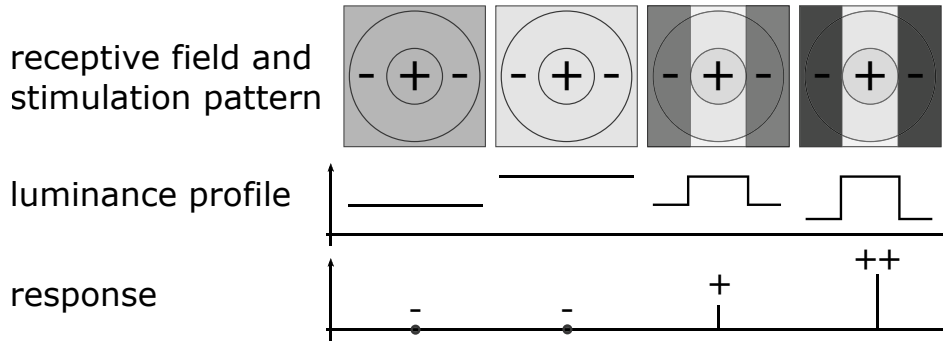

Fig. 2.5 Retinal receptive fields. On-center ganglion cells cover a receptive field with a central excitatory area (+) and a peripheral inhibitory area (-). The size of the receptive field varies across the retina. The luminance profile and the resulting response of the ganglion cell is shown for different patterns stimulating the receptive field. *Image after Ferwerda* [FPSG96]

Low-level Vision

The signals from the retinal photoreceptor cells are accumulated by connected collector cells and then transported to the retinal ganglion cells. These cells define a *receptive field* which allows the cell to receive signals from multiple connected collector cells. The size of the receptive field correlates to the size of a stimulus to which it is most sensitive. Light hitting the retina outside of the respective receptive field has no effect on the response of that cell. In the fovea each cone has a ”private line” to a ganglion cell. Otherwise the advantage of having a dense packing of cones would be lost. In contrast, towards the periphery, multiple cones or rods are connected to a single ganglion cell limiting peripheral vision resolution [AKLA11]. Additionally, neighboring ganglion cells may receive input from the same cone implying that receptive fields can overlap [Thi89].

The retinal ganglion cells always consist of an ON-region and an OFF-region in a concentric pattern. Depending on the response in correspondence to the center region, a ganglion cell is referred to as an ON-center cell or an OFF-center cell. The response of the cell is an antagonistic reaction between the surround region and the center region (Fig. 2.5). This unique property enables retinal ganglion cells to perform as high-pass image filter that is not direction-sensitive (*spatial tuning*) [Lue03].

In addition, ganglion cells are different in another functional aspect: The majority of ganglion cells (80%) are *midget ganglion cells* which are connected to the “small” parvocellular neurons (P-cells) in the visual pathways and transport primarily mid to high spatial and lower temporal frequencies, as well as red/green color differences. Second, the *small bi-stratified ganglion cells* (10%) carry blue/yellow color information for moderate spatial and temporal frequencies. Third, *parasol ganglion cells* (10%) are connected to “large” magnocellular neurons (M-cells) and are biased towards lower spatial but mid-temporal frequencies and deliver achromatic signals [Gol09, p.869]. The output of the retinal ganglion cells is transported via the visual pathways to the *visual cortex* (V1-V5 also known as Brodmann area 17-19) which is located in the occipital lobe of the brain. Midget cells

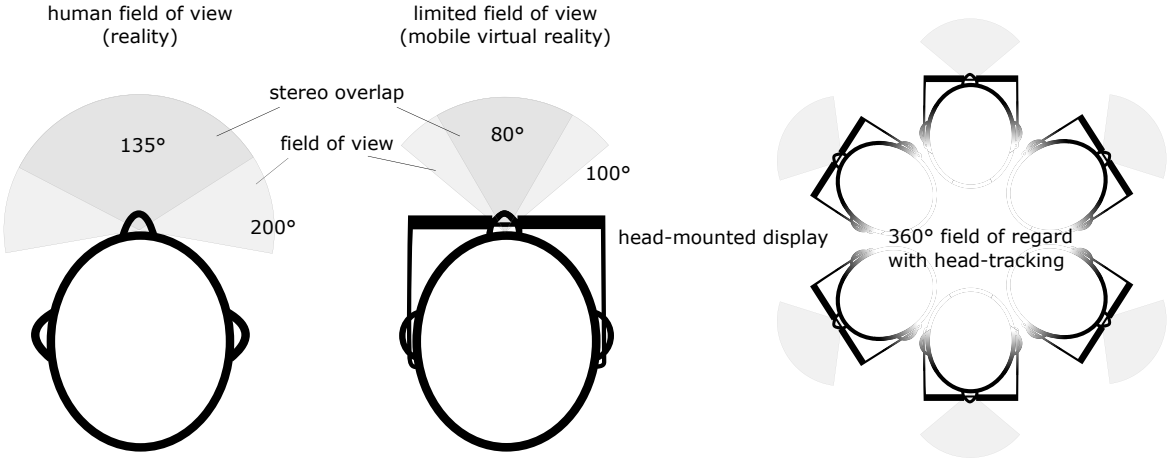

Fig. 2.6 Natural and constrained field of view. (left) With two eyes looking in the same direction human vision naturally achieves a wide field of view . (center) Common head-mounted displays significantly reduce the available field of view which may result in motion sickness and reduced depth perception. (right) With head-tracking sensors, the user can rotate his head resulting in a full field of regard of 360°.

transport signals to the primary visual cortex V1 whereas parasol cells transport to extrastriate cortical area V2. Similar to the ganglion cells, also cortical cells of the visual cortex have a receptive field as input. Two cell types exist in V1, both responding to contrast gradients: *Simple cells* respond to stationary or slow-moving stimuli and are orientation selective with a sensitivity of approx. 15 degrees (*orientation tuning*). *Complex cells* respond to moving stimuli and are selective to particular movement directions [Gol09].

Binocular Vision

Humans have binocular vision which means that we have two eyes pointing in the same direction with an overlapping field of view (FOV) from which visual information is fused into a single viewing area. Each eye has a FOV of 160° in horizontal and 135° in vertical direction, respectively. From the information of both eyes, the brain creates a combined field of view of 200°× 135°(Fig. 2.6, left). Depth perception is possible in the overlapping region of 120°× 135° [Wan95]. For centuries, researchers could not explain how the visual system produced an impression of a single world from both eyes. In 1838 Sir Charles Wheatstone presented the “stereoscope”, a mirror-based box that presents separate images to both eyes. This simple display enabled scientists to precisely manipulate the input to the binocular visual system and therefore to examine fusion of both eyes into a single view. Since visual information from each eye is slightly different our brain may interpret the spatial object differences (*disparity*) as depth so that the world is perceived three-dimensional. Owing to the low distance of the eyes, the depth perception disparity works best for close objects and is based on

other higher-level mechanisms for distant objects beyond 3m such as size, linear perspective, shading and interposition [Gol09, p.913].

Displays have evolved over time, and current head-mounted displays (HMD) provide stereo vision into convincing three-dimensional worlds rendered in real-time with a full field of regard (Fig. 2.6, right). However, due to limitations in the optics design most common HMDs significantly reduce the available FOV (Fig. 2.6, center). In addition, to enable depth perception, HMDs still follow the basic principle of the stereoscope by rendering a pair of images with a distinct viewpoint per eye. Unfortunately, in this case all objects in a scene are shown at the same focal distance due to the fact that they are shown on one single screen. Regardless of the virtual distance of the objects the eyes of the viewer constantly adjust focus so that the objects are clearly visible (*accommodation* and *convergence*). The difference between focus distance and virtual object distance results in conflicting depth cues for the brain. Using HMDs over an extended period of time can therefore lead to eye strain and an effect known as *motion sickness* arising by the "tunnel vision" from the reduced FOV and the conflict between accommodation and convergence (see Sec. 2.4).

2.3 Visual Sensitivity

Sensitivity of the HVS to visual detail varies spatially and temporally. Researchers are still not sure why the HVS does not provide uniform resolution. Currently, the best answer is that this would consume too many resources. Partly, this observation can be derived from physiological properties described in the previous section such as the physical acuity limit. Other explanations are based on empirical measurements of *visual acuity*, *cortical magnification* and the *contrast sensitivity function* which are discussed in this section. Importantly, these concepts are approximations of the HVS and still subject of active research. Many questions remain unanswered, such as the influence of attentional adaptation on visual sensitivity. The interested reader is referred to additional literature presented in the summary section.

The Foveal Spotlight

From the distribution of the retinal photoreceptors, one can expect that the spatial resolution is non-uniform over the field of view from both eyes. Indeed, we receive much more visual detail from the central viewing area than from the periphery. However, under normal conditions we do not have the impression of a non-uniform spatial resolution. This perceptual effect is known as the *foveal spotlight*. Under normal conditions when we *look at* things, our eyes are oriented towards the object without effort. This happens in a way that the area of interest is projected onto the region of highest spatial resolution, the fovea, so that it is viewed with the high-resolution spotlight. While reading this text the direction of the high-detail spotlight is wandering across the text. The HVS fuses the spotlight information with the surrounding low-detail peripheral data into one, coherent, seemingly high-resolution mental representation [Gol09, p.454].

Visual Acuity

"Visual acuity is a measurement of the keenness of sight" [AKLA11]. Historically, this property of human vision was measured by the minimal angular distance of two stars that can still be distinguished. More precisely, acuity can be described in the following ways:

- Minimum visible acuity - detection of a feature
- Minimum resolvable acuity - resolution of two features
- Minimum recognizable acuity - identification of a feature
- Minimum discriminable auity - discrimination of a change in a feature

The *minimum visible acuity* provides an estimate of the smallest visual detail the HVS is spatially able to resolve. A good example under ideal conditions could be a dark wire in front of the bright blue sky. Since the optics of the eye is not perfect, it spreads the image of the wire onto a wider area of the retina. Hence, a row of cones receives less light. Under ideal conditions the minimum visible acuity achieves an angle of just 0.5 arc seconds ($\approx 0.00014^\circ$) [HM39].

The *resolvable acuity* (also *acuity limit*) provides a fundamental limit on spatial vision that can be defined as the "finest distinction between two high contrast features". The acuity limit can be derived from the spacing of photoreceptors in the retina: In the fovea, the most densely packed region of the retina, cones subtend around 0.5 minutes of visual arc [TFCRS11].[4] The cone spacing of 0.5 arc minutes results in a spatial period of about 1 minute. From the *Nyquist limit* it follows that this period corresponds to a grating spatial frequency of 60 cycles per degree (cpd) since two cones per cycle, each connected to a separate ganglion cell, are required to resolve the spatial frequency. Towards the periphery visual acuity gets lower since multiple photoreceptors are connected to one ganglion cell and the packing of photoreceptors is less dense. It is known from sampling theory that *aliasing* occurs if a signal contains frequencies higher than the observer's Nyquist frequency [Sha49]. In human vision, this undersampling effect occurs if spatial frequencies higher than approx. 60 cpd are received in the retina. The aliasing in human vision has been demonstrated by projecting a projected high-frequency pattern onto the retina [Wil85]. Usage of a laser interferometer for projection enables to bypass the blurring effect of the cornea and lens so that the tested person actually perceives an aliased pattern. Fortunately, the optics of the eye diminishes the contrast of high spatial frequencies at the wavelength-dependent cutoff spatial frequency so that aliasing is not perceived. The cutoff frequency f in cpd can be estimated by the equation $f = \pi/180 \cdot p_d/\lambda$, where p_d is the pupil diameter in mm and λ is the wavelength of light in mm [AKLA11]. As an example, the cutoff frequency for a 3 mm pupil and a wavelength $\lambda = 600nm$ is 87.27 cyc/deg. Hence, for plausible pupil sizes of 2 to 8 mm, the resulting cutoff frequency exceeds the perceivable high frequencies.

The *minimum recognizable acuity* describes the angular size of the smallest feature that can be identified. This size, as well as the minimum resolvable acuity, can be estimated by the famous *Snellen*

[4] The "rules of thumb": 1.5°, 2.0°, and 8-10° visual arc correspond to roughly the apparent sizes of the thumbnail, thumb joint and fist at arm's length. 1°= 60 minutes of arc.

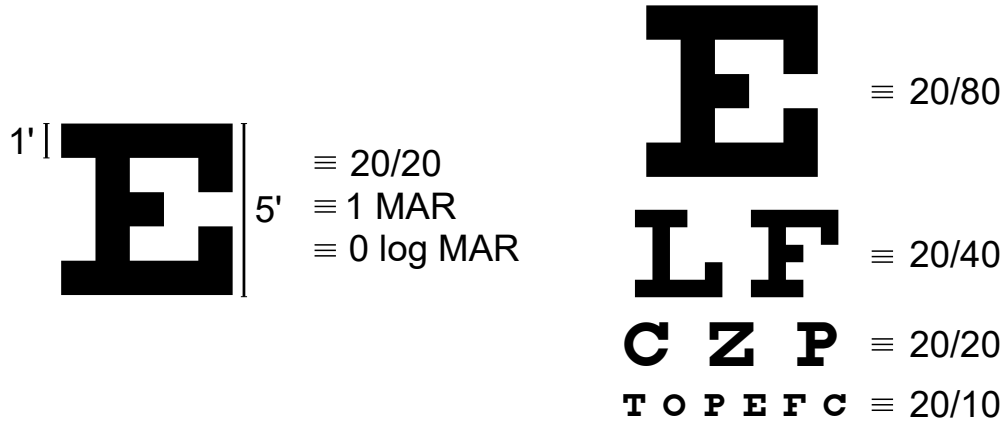

Fig. 2.7 Snellen letter and chart. Letters of specific sizes and line spacing form the basis for traditional acuity testing. Common sizing conventions are given (Snellen values, MAR, log MAR).

test. In this test letters of defined sizes are presented to the subject at a distance of 20 feet. The task is to identify the letter on the Snellen chart correctly. For example, the spacing between the bars of the letter "E" may be exactly 1 arc minute, and the entire letter is 5 arc minutes high. Normal vision can be confirmed in case the tested person is able to identify the letter correctly (20/20). However, most healthy young adults have better acuity (20/15). The size can also be expressed as *minimum angle of resolution* (MAR).[5] Normal vision corresponds to 1 MAR. If smaller letters can be identified, vision is better than normal (e.g. 20/10 or 0.5 MAR). The acuity limit is lower if only larger letters can be identified on the Snellen chart (e.g. 20/30, or 1.25 MAR).

Visual Acuity depends on the contrast of the test image or letter. Therefore, the acuity limit is measured for high contrast and under photopic luminance conditions, which corresponds to typical daylight and display use cases (80-320 cd/m^2). Visual acuity reduces significantly for mesopic and scotopic luminance levels (20/200) [AKLA11].

Reduced vision performance may have different individual reasons: myopia, commonly known as near-sightedness, or an eye disease, or a reduced refraction quality of the lens due to aging. In many cases, glasses can compensate for reduced refraction performance so that vision quality again achieves 20/20 Snellen and is considered as *corrected-to-normal*[6].

Last but not least, the *minimum discriminable acuity* describes the angular size of the smallest change in a feature that a person can discriminate [AKLA11]. This change can be in size, position or orientation. The smallest misalignment of two horizontal line segments that the HVS can discern is known as *Vernier acuity* [LKA85]. Measured by this definition, human vision provides an acuity limit of just three arc seconds ($\approx 0.0008°$) which is one magnitude higher than what is expected from

[5]Depending on the application also the logarithm of the minimum angle of resolution, $\log_{10}$ of MAR, is a popular unit [AKLA11].

[6]Rule of thumb: 1 diopter of uncorrected myopia results in a decrease of Snellen acuity to ≈20/60.

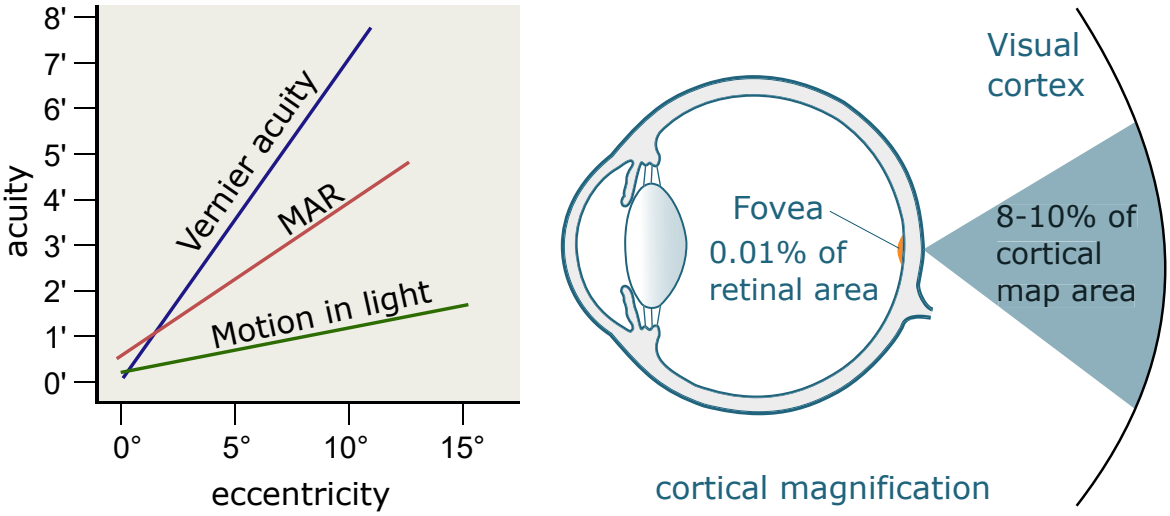

Fig. 2.8 Visual tasks and cortical magnification. (left) Different acuity measures can be described by linear functions with respect to eccentricity, such as minimum angle of resolution (MAR), line segment disalignment (Vernier acuity), or the minimal perceivable positional change of a feature (Motion in light). (right) The cortical magnification maps the small area of the fovea to a much larger area on the visual cortex. *Image after Weymouth* [Wey63] *and Goldstein* [Gol09]

cone spacing. This observation and other related tasks have created the term *hyperacuity*. However, hyperacuity does not refer to a greater ability to resolve fine detail.

Acuity Models

Due to the non-uniform distribution of photoreceptors in the retina the minimum resolvable acuity varies with eccentricity, i.e., the distance to the fovea given in degrees of visual arc. However, studies have shown that at eccentricities greater than two degrees actual acuity differs from what would be expected from average cone spacing [Gre70].

The psychophysical model by Aubert and Foerster from 1857 is the first model that mathematically describes visual acuity distribution [AF57]. Based on letter acuity measurements, Aubert and Foerster derived a map of isopters, i.e. lines of equal acuity. Interestingly, the isopters have elliptic shapes resulting in a rather anisotropic acuity distribution. The model follows approximately the distribution of cones and rods in the retina. It reaches its highest value in the foveal region. In the periphery acuity falls-off rapidly with eccentricity.

Later, Weymouth has shown that many visual tasks described as functions of eccentricity decrease roughly linearly with eccentricity for the first $20-30$ degrees [Wey63]. Visual performance decreases more rapidly for higher eccentricities [Wey63]. His results for visual acuity, i.e. the *minimum angle of resolution* (MAR), vernier acuity, and the minimal perceivable motion of a small stimulus are shown in Fig. 2.8 (left).

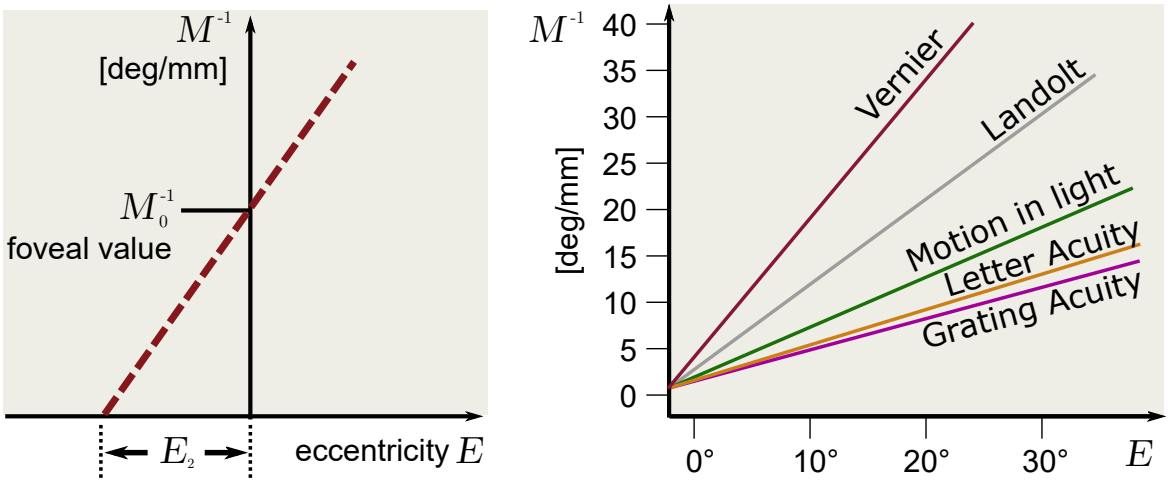

Fig. 2.9 Cortical magnification definition. Levi's E_2-value and the inverse foveal value M_0^{-1} (left) allow for a linear description of the inverse cortical magnification for different visual tasks (right). *Image after Strasburger* [SRJ11]

Cortical Magnification

Levi and collaborators parameterized visual performance across the visual field by a gradient called E_2 and a *foveal value* M_0 [LKA85]. The E_2 value represents the eccentricity at which the task-specific foveal value has doubled (Fig. 2.9, left). An explanation for the linear behavior has been provided through the concept of *cortical magnification* by Whitteridge & Daniel and Cowey & Rolls [DW61, CR74]. The cortical magnification factor (CMF) M represents a mapping from visual angle to a cortical diameter in mm (Fig. 2.8, right). The factor M is largest for those areas corresponding to the fovea and decreases with eccentricity for "peripheral" areas. In the fovea 1 degree viewing angle is mapped to a cortical distance of 20 mm. At 10 degrees eccentricity the cortical distance has reduced already to 1.5 mm. The CMF is usually given in mm/deg but plotted inversely (deg/mm). The CMF describes neuroanatomical properties. However, it has been shown for many acuity tasks that M can be *directly* measured by psychophysical approaches [DW61]. The inverse CMF (Fig. 2.9) is described by the following linear function [RV79]:

$$M^{-1} = M_0^{-1} \cdot (1 + E/E_2), \text{where} \tag{2.1}$$

- M^{-1} is the inverse cortical magnification factor,
- M_0^{-1} is the task-specific *inverse foveal value*,
- E is the eccentricity in visual angle,
- and E_2 is the eccentricity at which magnification has fallen by a factor of 2.

One striking advantage of the inverse cortical magnification concept is the ability to establish comparability of different visual tasks. Figure 2.9 (right) shows the CMF for Vernier acuity ($E_2 = 0.64$), Landolt rings acuity ($E_2 = 1.0$), differential motion in light ($E_2 = 1.75$), Letter acuity ($E_2 = 2.32$) and Grating acuity ($E_2 = 2.63$) normalized for the foveal value $M_0 = 1$ [SRJ11].[7] Comparison of different slopes requires both the gradient E_2 and the foveal value M_0.

The applicability of the E_2 value to many different visual tasks gave evidence that the variation of performance across the visual field is related to the mapping properties of the visual pathways. Resulting from the linear cortical magnification the M-scaling hypothesis has been derived which claims that performance degradation with eccentricity can by canceled out by scaling stimuli spatially. For example, in order to compensate for the loss in acuity when trying to read letters in the periphery, those letters just have to be enlarged in accordance to the linear CMF (Fig. 2.9, right) to be equally readable again. This method has been successfully demonstrated by Cowey and Rolls [CR74] and motivated researchers to unify fovea and periphery. Strong supporters of the concept claimed that "a picture can be made equally visible at any eccentricity by scaling its size by the magnification factor" [RV79].

However, other researchers have pointed out difficulties of the M-scaling concept [WM78]: First, the linear CMF model only approximates the complexity of the HVS, as peripheral vision is not a scaled-down version of foveal vision [BKM05]. Second, several studies exist in which the CMF concept is less convincing or clearly fails, such as stereo acuity, two-point separation in the far periphery, or contrast sensitivity for scotopic vision [SRJ11]. In addition, due to variations in the measurements for different visual tasks as well as due to inter-individual differences, it is still an open question whether M^{-1} is linear also at near-foveal eccentricities [SRJ11].

Acuity is affected by eye adaptation in very bright and dark areas. Additionally, eye motion and cognitive factors influence the amount of detail perceived [Gol09]. Therefore, acuity slopes over eccentricity (for simplicity described as *acuity* in the remainder of this thesis) are strongly task- and user-dependent and cannot be precisely predicted beforehand with any known model. However, Levi's E_2 and the CMF concept is still able to describe a large portion of performance variations across the visual field for many gaze-contingent visual tasks. The estimated E_2 values can therefore still be used as a yardstick.[8]

[7] A more comprehensive list of E_2 values and a list of failure cases for the CMF concept can be found in [SRJ11, p.12-15].

[8] A study with seven spatial threshold tasks by Virsu et al. shows that 85-97% of the variance was removed by M-scaling [VNO87].

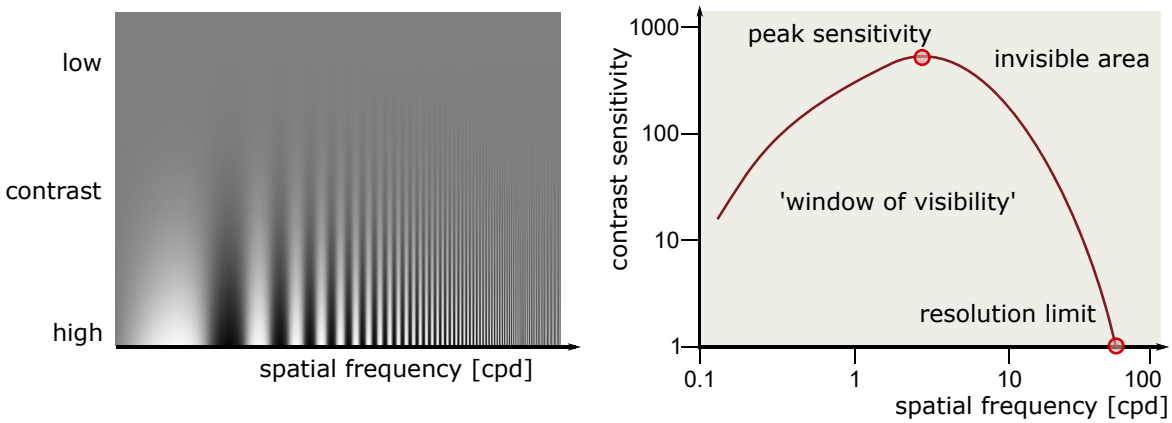

Fig. 2.10 Spatial contrast sensitivity function. (left) Sine wave pattern for estimating spatial contrast sensitivity (Campbell-Robson chart). Spatial frequency is modulated horizontally whereas contrast varies vertically. (right) The derived contrast sensitivity function (CSF) mirrors the inverted U-shaped region in which the wave pattern is visible. *Image after Snowden* [SSTT12]

Spatial Contrast Sensitivity

Spatial detail in a pattern can only be perceived if the pattern has sufficient *contrast*. Consequently, perceived spatial detail not only depends on the spatial frequency of a certain pattern but also on the amplitude of the pattern frequency [AKLA11].

One common variant to measure perceivable contrast are sine wave patterns of changing black and white stripes whereby spatial frequency increases from left to right and contrast increases from top to bottom (Fig. 2.10, left). One period in the sinusoidal grating at the projected size of 1 degree is defined as 1 *cycle per degree* (cpd). The higher the number of line pairs per degree viewing angle, the higher the spatial frequency in cycles per degree.[9] The sensitivity of contrast is measured by determining the lowest contrast for which the wave pattern is still visible. Usage of a sine wave pattern for contrast sensitivity estimation is reasonable since there are cells in the visual cortex that selectively respond to contrast and different spatial frequencies [Lue03].

The contrast sensitivity function (CSF) is derived by computing the inverse of the contrast required for threshold detection (Fig. 2.10, right). The region under the curve is commonly coined *the window of visibility* [AKLA11]. The resolvable acuity limit (60 cpd) corresponds to the lowest contrast sensitivity value. Contrast sensitivity peaks at about two cycles per degree (half the width of a finger nail at arm's length) and decreases to both sides towards higher as well as lower spatial frequencies. Very high (>60 cpd) and very low frequencies (<0.1 cpd) cannot be perceived at all. The upper limit corresponds to the acuity limit derived from cone spacing in the last section. However, the lower limit cannot be directly derived by the eye's physiology. Recent research has given evidence that limits for

[9] Snellen values can be converted to cpd by multiplication of the Snellen denominator with 30: for example 20/20 converts to 600/20 = 30 cpd. 20/200 converts into 600/200 = 3 cpd.

contrast sensitivity may be explained by a combination of optical and neural properties of the HVS. One explanation could be that for lower frequencies more rods must be not firing due to the receptive field size. Due to measurable noise in rod activation, this state gets more unlikely with decreasing frequency [AKLA11].

From the fovea to the periphery, contrast sensitivity decreases significantly at all frequencies but fastest for high frequencies [RVN78]. As for visual acuity, *M*-scaling can approximately compensate for the performance reduction. Rovamo et al. achieves comparable peak sensitivity values across the visual field if the stimulus size is scaled by the receptive field at each eccentricity, a factor derived from retinal ganglion cell density. Hence, contrast sensitivity at any eccentricity depends on the number of ganglion cells stimulated by the respective grating pattern [RVN78]. A CSF model including spatial frequency and retinal velocity is proposed by Daly et al. [Dal98].

In addition, contrast sensitivity depends on the chromaticity of the stimulus[10]. Previous measurements focused on achromatic light. However, the fovea is tuned to chromatic red/green stimuli, whereas those stimuli are significantly less salient in the periphery. Blue/yellow and achromatic stimuli result in a less pronounced decrease in terms of contrast threshold [Mul85].

Adaptation

A circumstance with which the human eye deals impressively well is the change of total luminance. Ranging from low light-levels at night to bright sunlight, basic visual tasks are maintained without effort. The human eye can cope with illumination levels between 10^{-4} and 10^{8} cd/m^2. The adaptation mechanisms to a lower or higher luminance can be grouped into pupil size adjustment, retinal photoreceptor adjustment, and neural activity adjustment in the visual cortex. However, retinal and neural adjustment contribute significantly more to the overall adaptation effect. Hence, *Adaptation* can be defined as the time-dependent process of tuning sensitivity of retinal cells and neurons to the amount of incoming light.

The pupil's contribution to dark and light adaptation takes only a few seconds to be completed. Triggered by the pupillary reflex the iris can adjust pupil diameter from 2 to 8 mm [Gol13]. Therefore, the pupil area can change by a factor of 16. Hence, only about one magnitude of light intensity difference (1 log unit) can be controlled by adjusting the pupil size. Presumably, with the change of pupil size the HVS primarily tries to limit the optical effect of aberrations [FPSG96]. Therefore, most of the adaption task takes place on the retina level.

Given that there are two different kinds of receptors present, rods and cones, adaptation needs to be addressed separately. In fact, the sensitivity of either system is not affected by the stimuli of the other one at all (*receptoral duplicity*) [HF86]. Additionally, sensitivity of neurons in the visual cortex adjust to different light levels during adaptation [AKLA11]. As a result, adaptation enables the HVS to perceive visual information robustly over seven orders of magnitude of brightness intensities.

[10] Midget ganglion cells (80% of total retinal population) carry red/green opponency information, bi-stratified cells (10%) carry blue/yellow information and parasol cells (10%) carry achromatic signals.

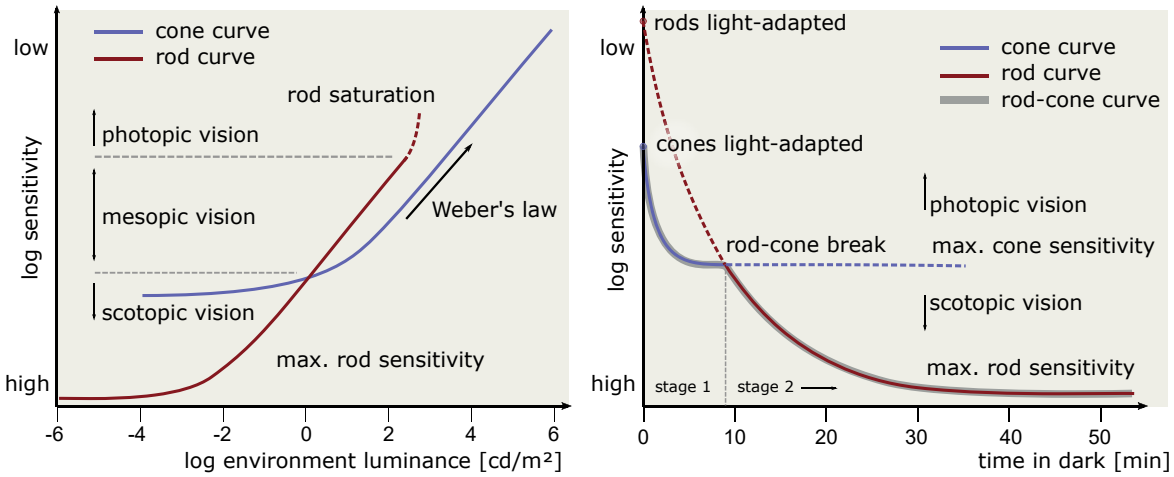

Fig. 2.11 Adaptation sensitivity curves. Light adaptation (left) and dark adaptation curves (right) for rods and cones. Active photoreceptors at each stage are visualized by a solid line whereas dashed line parts represent photoreceptor saturation or inactivity. *Image after Goldstein* [Gol13]

However, we are not able to see equally well at all intensity levels. It is therefore difficult to read a book in a dim room without additional light. Adaptation comes at the expense of reduced acuity at lower light levels. At daytime, contrast sensitivity is lower but visual acuity and color vision excel.

Almost everybody has experienced the situation when leaving a building and being blinded by the sunlight outside at daytime. Vision is significantly impaired and can be even painful at first, but within seconds our eyes get used to the sunlight. In perception literature this process is known as *light adaptation*. It reduces sensitivity of the HVS as light intensity increases. Conversely, *Dark Adaptation* describes the change of vision from brightness to darkness. For example when entering a cave or dark tunnel after having been in the sunlight, vision is temporally almost disabled but restored in a matter of minutes.

The gradient of inverse sensitivity versus environment luminance during light adaptation is shown in Fig. 2.11 (left). For very low luminance values, rods are most sensitive. For higher illumination values, visibility threshold increases proportionally to the square root of background luminance, culminating in a approx. linear increase. The linear gradient shows the effect of contrast constancy known as *Weber's law* [FPSG96]. In other words, with adaptation high sensitivity to contrast is restored. This is achieved by decreasing the intensity signals in the visual system by the same fraction. Hence, ratios among the photoreceptor signals remain relatively unchanged. For rods this behavior continues with increasing luminance until sensitivity is compressed and *saturation* is reached (dashed line in Fig. 2.11, left). Saturation can be described as the level of illumination where even low intensities result in full photoreceptor stimulation, preventing the HVS from distinguishing luminance differences. Importantly, the cone system does not saturate. [FPSG96].

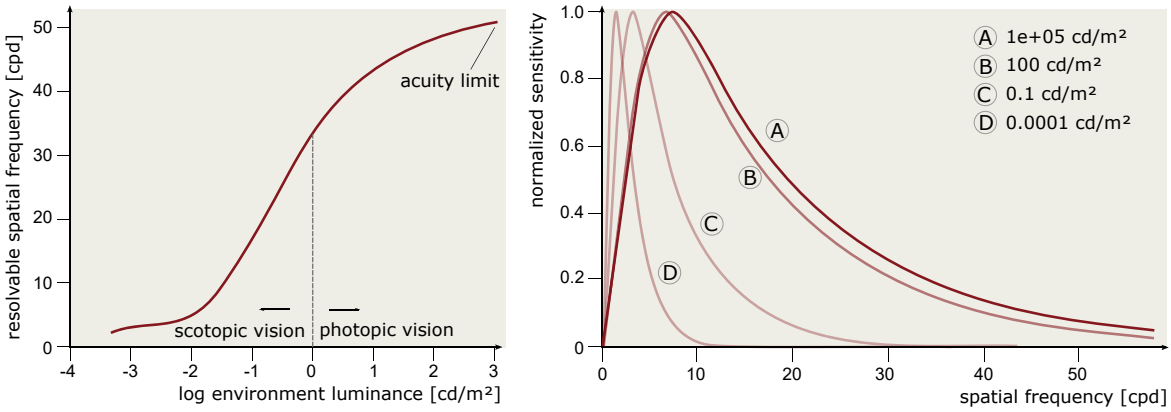

Fig. 2.12 Adaptation-dependent acuity and CSF. Spatial acuity increases non-linearly from scotopic to photopic vision (left). The normalized contrast sensitivity functions (CSF) is compressed with lower adaptation levels (right). *Image after Ferwerder* [FPSG96]

Temporally, the process of light adaptation happens very rapidly. At scotopic levels more than 75% of sensitivity is recovered within 200 milliseconds and the remaining amount in a matter of seconds [Ade82]. For photopic levels light adaptation occurs over a period of about 5 minutes [Bak49]. In contrast to light adaptation, rods during dark adaptation may require 35 minutes to reach highest sensitivity. The prolonged behavior of dark adaptation is visualized in Fig. 2.11 with the inverse sensitivity on the y-axis and the time in minutes on the x-axis. The initial sensitivity is given for photopic vision by the cone sensitivity at light-adaptation (time=0). Two distinct stages are apparent. In the first stage, known as *foveal adaptation*, sensitivity is increased for cones and rods. This process happens more quickly for cones (≈8 minutes) and increases sensitivity by about 1.5–2 log units. As the rod-cone break sensitivity for cones cannot be increased anymore, rod cells get more sensitive than cones. At this point perception changes from *photopic vision* using cone cells to *scotopic vision* using rod cells. The second stage (rod adaptation) is more protracted (20-30 minutes) and increases sensitivity of rod cells. The second stage results in a sensitivity change of another 4 log units. Importantly, dark adaptation is affected by the level of pre-adaptation [KNFJ14]. An increasing pre-adaption illumination results in a longer cone branch and delays scotopic vision. Moreover, the time to reach absolute sensitivity also increases.

Adaptation influences the performance of the HVS, such as color perception, spatiotemporal contrast sensitivity and the amount of perceivable detail [LSC04]. At scotopic levels absolute sensitivity is high, but since rods provide achromatic signals only, colors cannot be perceived [FPSG96]. In contrast, at photopic levels sensitivity is dramatically reduced but colors can be perceived due to the trichromatic nature of cone cells. The reduction of perceivable spatial detail with dimming of light is visualized in Fig. 2.12. The highest perceivable spatial frequency of a grating pattern reduces from 50 cpd at photopic levels (≈ 20/10 Snellen) down to 2 cpd for scotopic vision (20/300 Snellen). When

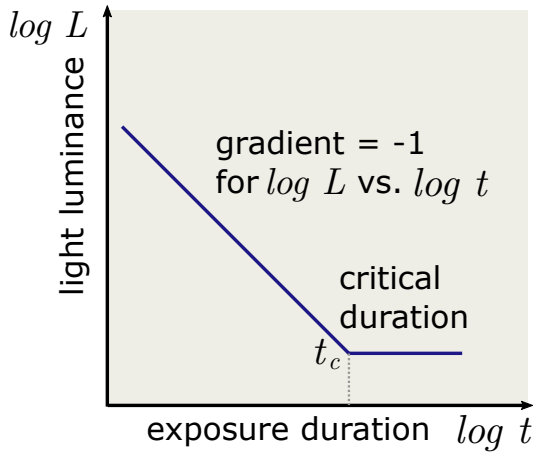

Fig. 2.13 Bloch's law. On a logarithmic scale the required amount of light to reach visibility depends linearly on the retinal exposure duration which is limited by the critical duration t_c.

Image after Adler [AKLA11]

visualized for all perceivable spatial frequencies lower adaptation levels result in a compression of the CSF (Fig. 2.12, right).

Temporal Sensitivity

In nature most temporal variation of light reaching the eye occurs through image motion induced by the observer, eye motion or object movement. Although we generally have the impression of perceiving the world continuously, basic temporal processing in the HVS has similarities to a camera taking discrete pictures of the object of interest. For taking a picture, the shutter of a camera is opened for a discrete duration of time during which light photons reach the light-sensitive camera sensor or material. Pixel intensity of the resulting image depends on the number of photons hitting the corresponding sensor pixel in the shutter interval (exposure) as well as on the sensitivity of the sensor (ISO value). Visual perception basically works in an analogous way in which exposure and sensor sensitivity can be, illustratively, substituted by features of the HVS, namely temporal summation and adaptation. *Temporal summation* is the process of collecting incoming photons by the photoreceptors in the retina and happens in a time interval of 20 to 30 ms (cones) being upper-bounded by the *critical duration* [AKLA11].

Temporal sensitivity cannot be studied in isolation since its performance varies across the visual field. Other stimulus properties, such as spatial dimension, color and background features, also influence the ability to perceive temporal changes. The remaining part of this section excludes scotopic vision and rod-specific models since low-light situations are less relevant for typical display settings. However, there is a rich psychophysical literature covering this area (see Sec. 2.6).

[11] Troland (named after Leonard T. Troland) is a unit of the retinal illuminance T : 1 Troland = 1 $\frac{cd}{\mathrm{m}^2} \cdot \mathrm{mm}^2$. The unit takes scaling by the pupil size into account: $T = L \times p_a$ where p_a is the pupil area in mm^2.

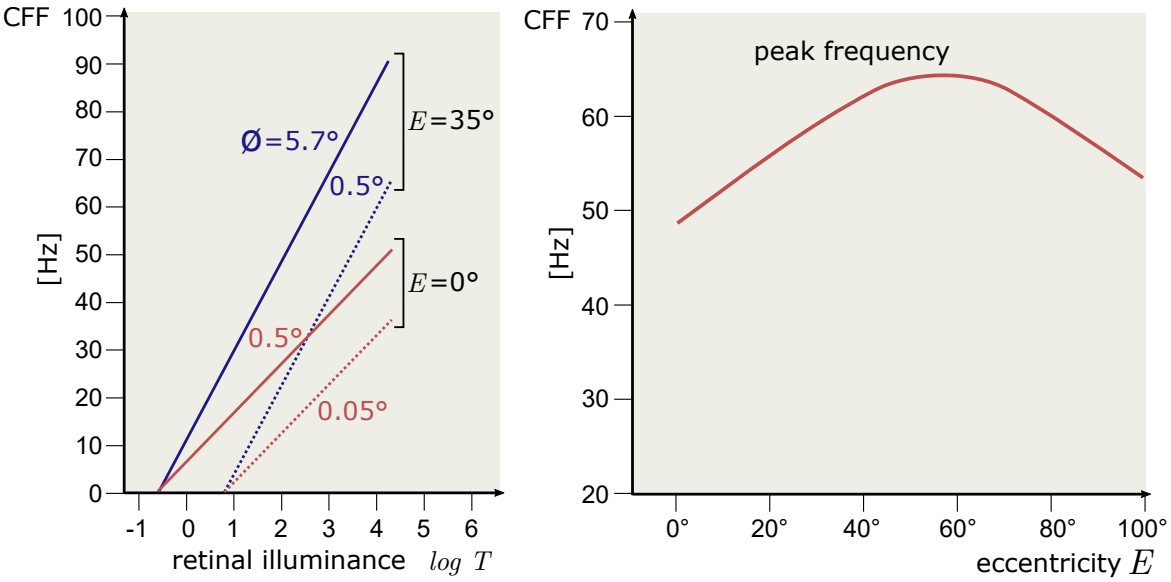

Fig. 2.14 Critical flicker frequency (CFF). (left) The CFF increases linearly with stimulus illumination (Ferry-Porter law) and with stimulus size (Granit-Harper law). The retinal illuminance T is given on a logarithmic scale in Trolands[11]. Stimulus diameters and eccentricities E are given in visual angle. (right) The CFF increases up to 55° eccentricity and decreases in the far periphery for a constant stimulus (stimulus area=88.4°, retinal illuminance T=2510 Td, pupil diameter = 8mm).
Image after Kelly et al. [Kel61]

The most basic question on temporal sensitivity is how much light is required so that a stimulus can be perceived. In order to reach visibility of aperiodic stimuli, a single pulse of light, the required exposure time t and luminance of the light L are coupled by the following equation known as *Bloch's Law*, $L\,t = k$, where the product of luminance L and stimulus duration t result in a constant value k (Fig. 2.13). When critical duration t_c is reached, the threshold luminance is constant for larger values of t since temporal summation cannot increase over the critical duration. Under experimental conditions the visibility threshold given by Bloch's law is valid across the full visual field and for a variety of background conditions [KNFJ14]. A stimulus presented for 30 ms at a luminance of 80 cd/m^2 is equally well perceivable as a pulse presented for 60 ms at 40 cd/m^2 [SHH07].

However, in a real-world setting the critical duration is harder to define since photoreceptor cells in the retina are affected by noise. This noise during the temporal summation interval of each cell results in more probabilistic sensitivity behavior known as *probability summation*. In addition, the critical duration value depends on a variety of attributes, such as adaptation level, spatial frequency, size, chromaticity and eccentricity of the stimulus. The critical duration t_c is longer for lower adaptation levels and for higher-frequency, smaller, chromatic stimuli at higher eccentricities [AKLA11].

Returning to the example of a camera taking a picture of a light source with defined luminance, Bloch's law would describe the *minimal* shutter time necessary so that the light source is *barely*

visible on the resulting picture. If one wants to record video instead of a single picture, the frame rate estimates how many frames are captured in a certain amount of time. For the HVS a related value can be given known as the *critical flicker frequency*: "The *critical flicker frequency* (CFF, also flicker fusion frequency) describes the fastest rate that a stimulus can flicker and just be perceived as a flickering rather than stable" [AKLA11, p.700].

Correspondingly, the CFF has many practical applications for the development of displays.If a light flickers faster than what the HVS is able to resolve we perceive the flashing light as stable rather than seeing a sequence of flashes. The CFF is dependent on different features. Most interestingly, for photopic lighting conditions the CFF increases linearly with log luminance of the flickering light over a dark background. This behavior is known as the *Ferry-Porter law* and holds for a wide range of eccentricities [Por02]. The *Granit-Harper law* states that the CFF increases linearly with size of the stimulus area [GH30]. Both mentioned principles can be seen in Fig. 2.14 (left) in which the CFF function is plotted against retinal illuminance for different stimulus sizes and at different eccentricities.

Rovamo and Raninen have shown that in case stimulus size and luminance are constant, the CFF increases with eccentricity up to 55° eccentricity (see Fig. 2.14, right) [RR88]. Towards the far periphery the CFF decreases again. Hence, mid-peripheral vision has better temporal resolution than foveal vision and the far-peripheral vision. This property of the HVS can also be observed when a traditional CRT screen (cathode ray tube) at 50 Hz frame refresh rate appears constantly illuminated in the foveal area but is perceived as flickering when viewed peripherally. If the CFF is plotted against the number of stimulated retinal ganglion cells, the resulting function is linear across all eccentricities [RR88].

Temporal Contrast Sensitivity

The critical flicker frequency has been introduced as the threshold frequency at which a periodically flickering light is being perceived as constant. The CFF curve is valid for maximum contrast and large size of the flicker stimulus only. In the following, temporal contrast sensitivity below the CFF and for stimuli of varying contrast are briefly discussed.

As has been pointed out for the Ferry-Porter law, the CFF depends on the retinal illumination. Figure 2.15 shows the estimated temporal sensitivity for different retinal illuminance values at photopic levels with an achromatic flickering stimulus. In this chart, temporal frequency along the x-axis is plotted against the modulation ratio of the flickering stimulus. The modulation ratio represents a percentage deviation of the amplitude of the stimulus from its average value. An modulation ratio of 1 results in CFF values as before whereas lower values result in less pronounced flickering. Hence, the higher the curves representing the threshold modulation ratio the higher temporal sensitivity for a given adaptation level [AKLA11]. It can be seen that at low frequencies modulation sensitivity is approximately equal for all adaptation levels. For higher flicker frequencies modulation sensitivity

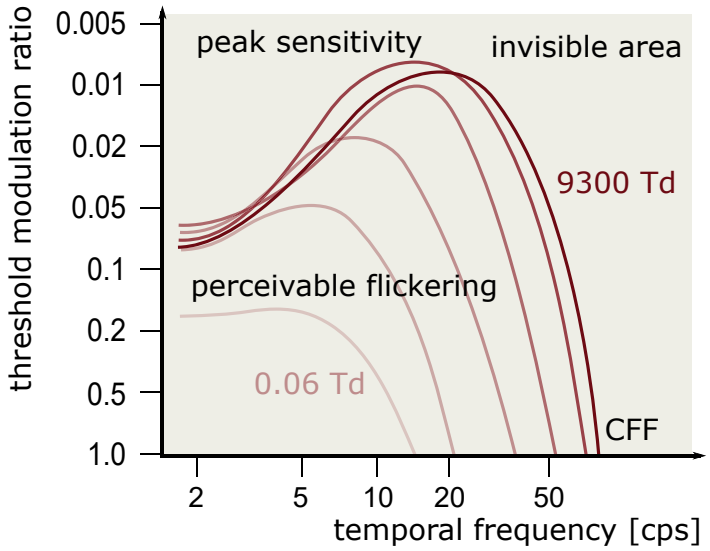

Fig. 2.15 Temporal contrast sensitivity function (CSF) for different retinal adaptation levels. Each curve represents the threshold modulation ratio (percentage deviation of average value) of a just detectable flicker stimulus for a given adaptation level (in Trolands) plotted against the flicker frequency (in cycles per second, cps). Low levels of retinal illuminance result in a low-pass CSF whereas higher levels reshape the CSF into a more band-pass curve. *Image after Adler* [AKLA11]

strongly depends on retinal illuminance values. Hence, highest sensitivity peaks at about 20 Hz only for high adaptation levels.

As for many other attributes of the HVS, the temporal CSF shown in Fig. 2.15 does not show the complete picture. Many other properties result in deviation from the shown CSF behavior. For example, chromatic flickering stimuli result in more low-pass CSF curves that are cut-off at lower temporal frequencies in comparison to achromatic stimuli [Kel61].

Interestingly, sensitivity is decreased if the flicker stimulus is presented on a background with large contrast for photopic conditions due to rod-cone interactions. This effect is pronounced for the periphery and for low temporal frequencies [CA84]. Hence, the HVS is most sensitive to temporal changes if the average luminance of the flicker stimulus matches the background luminance [AKLA11]. Further, the perceived brightness of a light as well as its perceived color can be altered by flickering. This effect is called *apparent brightness* and has been recently exploited to improve perceived color saturation of images beyond the display capabilities [MFN16]. Kelly et al. explored spatial contrast sensitivity in combination with temporal contrast sensitivity. As a result, the spatiotemporal contrast sensitivity function in [Kel61] provides modulation sensitivity values for a given pair of temporal and spatial frequencies of a stimulus. However, the shape of the resulting surface is not completely understood. Explaining these findings is still a topic of active research.

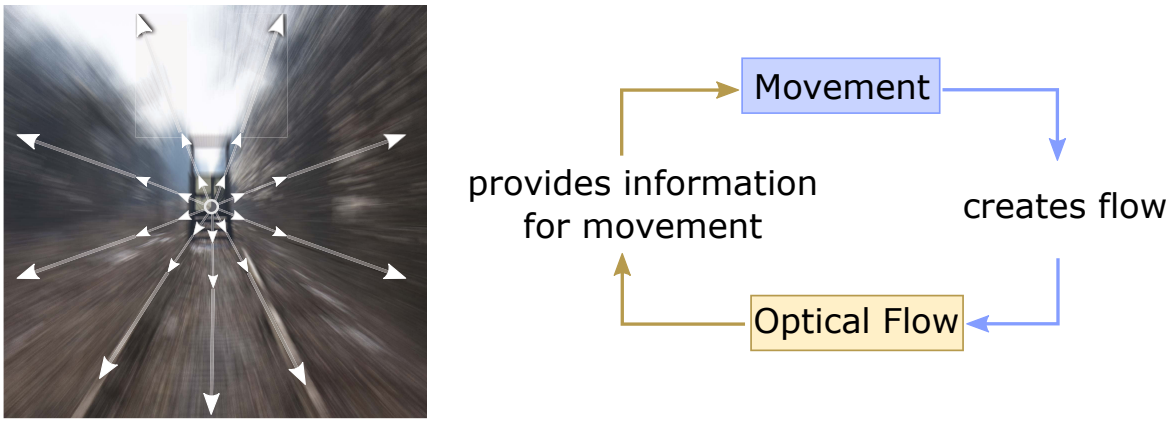

Fig. 2.16 Optical flow pattern. (left) When an observer is approaching a target head-on, characteristic radial flow patterns arise. There is almost no flow in the area of the target. The flow increases radially with distance from the target. (right) The HVS constantly uses flow patterns to filter object motion from self-motion. *Image after Goldstein* [Gol09]

Motion Processing and Optical Flow

The motion of an object in the visual field results in a change of its projection on the retina. The HVS is extremely sensitive to motion. Even a small change at any place of the full visual field can immediately grasp attention. This property of the HVS is critical for many daily tasks, e.g., obstacle avoidance or driving. Interestingly, the minimum shift at which a movement can be detected is in the range of Vernier acuity for photopic conditions ($\approx$ 20 arc seconds, see Fig. 2.8) [Bas06]. This minimum shift threshold anisotropically increases with eccentricity. As it has been observed for isopters of visual acuity also motion perception decreases more quickly in the vertical direction when compared to the horizontal direction. This increase can be approximated by magnification theory [MN84].

Aubert and later Basler demonstrated that motion sensitivity not only depends on the extent of the movement but also on the velocity. A moving dot was easier to detect when motion velocity increased [Bas06]. Basler recognized that motion sensitivity also depends on the contrast. Hence, very small motions are perceivable within a bright lighting situation but are imperceptible with less lighting. The direction of the motion had no influence in the experiments.

Many effects of motion processing are explained by special neurons with directionally selective receptive fields which have been found in the visual cortex (V1) [AKLA11]. The neurons are especially sensitive to motion gradients. Therefore, an object moving relative to its background seems to pop out of the environment. In addition, strong motion gradients are perceived for occluding and revealing objects when the observer moves laterally (*parallax*). From this shift the HVS derives valuable information about the relative distance of both objects and the distance from the observer in the line of sight [Gol09]. The adaptation behavior of direction-selective neurons also explain the

known *motion aftereffect*. After prolonged perception of a moving object the stationary environment or slower-moving objects seem to be moving into the opposite direction. The velocity of objects is often miscalculated in this situation [AKLA11].

To perceive an object as moving our HVS must filter the retinal motion produced by the object from the retinal motion that is due to head and eye movements. The visual system also includes information from the vestibular system for disentangling object motion from self-motion. This complicated process makes use of an effect known as the optical flow.Optic flow is the whole-field retinal image motion of the visual field produced by self-motion and object motion.

For example, as an observer moves forward, the light reflected by the environment and received by his retinal receptors appears to flow past him in backward direction. Characteristically, the flow is faster for objects close to the moving observer. The flow is zero towards for the point the observer is moving (Fig. 2.16). The HVS exploits optical flow for a variety of tasks such as movement correction, velocity estimation, depth perception and time-to-contact approximation.[12] The concept and effects of optical flow has been the subject of research in psychophysics and computer vision for decades [LK80].

The optical flow can create the illusion of motion in case retinal object motion mimics characteristic flow patterns and the result does not conflict with the vestibular system. For example, when sitting in a departing train and looking outside it may feel for the observing person as if the world outside begins to move instead of the slowly accelerating train. This is known as the *aperture problem*.

In Virtual Reality, contradictory signals from retinal motion and self-motion can result in an uncomfortable psychophysical state known as *motion sickness*. This effect can only be avoided if potentially contradictory signals are reduced to a minimum and, in addition, by reducing the lag time (*latency*) between user movements and the system's response [SC02].

2.4 Eye Motion

Our eyes are constantly in motion. Six external muscles allow precise and fast changes of the horizontal and vertical orientation of the eye, independently from head orientation. The primary goal of moving the eyes is to move the projection of the object of interest (OOI) onto both foveæ so that the object is perceived with high detail. This mechanism allows exploration and scanning of the environment, shifting attention from one object to another. In addition, the eye muscles allow adjustment of the eye's lens to set the OOI into focus. The performed eye movements are briefly discussed in the following, namely saccades, the vestibular-ocular reflex, and smooth pursuit tracking.

Stabilization reflex during head movements During head movements, such as walking, the HVS uses acceleration information from the vestibular system as well as information of the amount of head rotation and retinal velocity information (optic flow) to keep the orientation of the eyes in

[12] Neurons that respond to optic flow patterns have been found in the *medial superior temporal area* (MST) [Gol09, p.429]

alignment with respect to the OOI. This *vestibular-ocular reflex* happens quickly with a latency of 7-15 milliseconds and is robust also for fast head movements [AKLA11].

Scanning the environment Humans constantly scan their environment sequentially. The most important mechanisms in this context are *saccades* and *fixations*. Saccades denote the motion when rapidly jumping from one object of interest to another one. Saccades can reach peak angular speeds of up to 900°/s [FR84], resulting in a dramatic decline in visual sensitivity during saccades referred to as *saccadic suppression*. Hence, during saccadic eye movements visual information cannot be acquired [WDW99]. In contrast, fixations describe the state and duration in which visual information is perceived while our gaze rests on an OOI. Fixation durations typically vary between 100 milliseconds and 1.5 seconds [WDW99]. It is assumed that the duration of a fixation corresponds to the relative importance and visual complexity of an area in the visual field. If more information needs to be processed, fixations typically take longer. When viewing a typical natural scene, the HVS triggers around 2–3 saccades per second [KFSW09] and the average fixation time is about 250 milliseconds. The spacing between fixations is, on average, around 7° viewing angle. Maintaining fixations at larger eccentricities (>30°) is uncomfortable and usually result in a head rotation towards the target, followed by a fixation at lower, more comfortable eccentricity.

Object tracking The unconsciously triggered tracking reflex when a moving object attracts our gaze is called *smooth pursuit eye motion* (SPEM). This eye motion enables the observer to track slow-moving targets so that the object is fixated onto the fovea. Interestingly, small eye movements up to 2.5 °/second have hardly any effect on visual acuity [AKLA11]. Researchers have found that the peak velocity for smooth pursuit eye motion is around 100 °/second [WDW99]. However, the success rate depends on the speed of the target and decreases significantly for angular velocities >30 °/second.

Compensatory Eye Motion While consciously fixating an object, the eye still performs tiny but important movements known as *tremor motion*. This unconscious motion refreshes the retinal image. Tests have shown that the perceived image fades away if tremor motion is inhibited [AKLA11].

Accommodation is the mechanical ability of the eye to change the shape of the lens so one can focus at different distances. When the ciliary muscles at the front of the eye tighten, the curvature of the lens and correspondingly its focusing power is increased. Accommodation describes the natural counterpart of adjusting a camera lens so that an object in the scene is set into focus. Importantly, this process happens unconsciously and without any effort in less than a second at photopic illumination levels [Gol09]. In addition to previously discussed visual cues, such as occlusion and parallax, accommodation is a strong clue of depth perception [HCOB10].

Vergence Motion This eye motion is coupled with the fixation process for binocular vision so that both eyes' gaze aims at the same point at a distance. Due to their positional difference both eyes

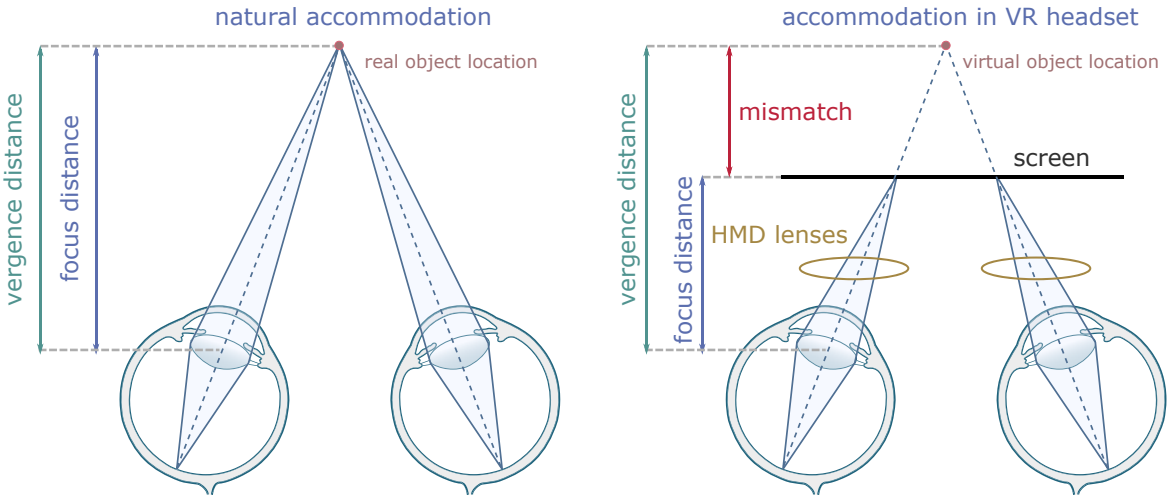

Fig. 2.17 Accommodation in reality and wearing a VR headset. (left) Accommodation and convergence are naturally both adjusted to the same distance. (right) In a VR headset static accommodation to the fixed-distance display is in conflict with the vergence distance of the displayed virtual object location.

receive the OOI from slightly different viewpoints. The difference of the per-eye gaze directions can be quite large when looking at an object close-by. *Vergence* moves the point of intersection of both gaze lines to the point of focus and allows humans to optimize the FOV overlap for a wide range of distances (Fig. 2.17, left). From the pair of stereo images, the HVS derives depth information resulting in three-dimensional perception in nature and in stereoscopic images.

Accommodation-Convergence Conflict In a VR headset the focus distance is fixed because the optical elements are static. In this case natural accommodation is not possible. With stereoscopic content this results in an accommodation-convergence conflict as the eyes converge to match the depth of the displayed virtual scene while the HVS has to accommodate for the – probably completely different – static focal distance (Fig. 2.17, right). As a consequence, over time this conflict can contribute to visual discomfort, fatigue, nausea, eyestrain and compromised image quality known as *motion sickness* [HGAB08].

2.5 Attentional Effects on Visual Perception

Why do we pay attention to certain areas of a scene but not to others? Do we have to directly look at an object to perceive it? Some answers to these questions are given in this section.

When our gaze is shifted from one object to another, the HVS is doing more than just *looking* at it. Attention is directed to specific attributes so that related features become more “clear and vivid” than unattended ones [Gol09]. As an example, Carrasco et al. have shown that for two identical grating

patterns contrast is perceived as being higher for the one to which the participant pays attention to [CLR04]. Attention affects how we see and experience and how well we perform a visual task. In a more practical situation, e.g. when waiting for a colleague at a café, we may pay attention to the door (where he or she is likely to appear) and attention to black objects (because he or she often wears black shirts). This is necessary since the limited processing power of the brain precludes evaluation of every incoming signal from the retina and visual pathways [Gol09]. The influence of attention on visual perception is known as the *attentional spotlight*. It is related to the foveal spotlight presented in Sec. 2.3. Without attention the perceived environment is seen as if it was a blurred image. Attention represents a spotlight that brings spatial locations into focus. Selective attention moves the attentional spotlight around in the image. Consequently, objects within the attentional spotlight are processed more accurately than unattended objects.

Directing gaze by eye movements, known as *overt attention*, is one important mechanism for selective attention. However, experiments have shown that we can also pay attention to things that are not directly in our central field of view, known as *covert attention* [Gol09]. In addition, we can look directly at an object without paying attention to it known as *inattentional blindness* or *cognitive tunneling*. Hence, visual attention and eye movement do not necessarily coincide or follow the same patterns [TW01].

What determines what we pay attention to and where we fixate our gaze in a scene? *Stimulus salience* refers to the visual "attractiveness" or importance of the environment. From a bottom-up point of view, the detection of objects across the visual field is assumed to be subconscious and does not depend on attention (pre-attentive processing) [WDW99]. According to this theory, the salience of a stimulus is affected by low-level features such as color, orientation, brightness and contrast of the stimulus. Researchers have successfully created bottom-up models to predict possible fixation locations in images and videos with high probability [IKN98]. However, the bottom-up perspective is neither sufficient for prediction of the actual sequence of fixations, known as the *scan path*, nor for fixation duration since selective attention is not just based on low-level features [OTCH03, HBCM07].

Recent research has revealed that attention is primarily driven by cognitive factors such as the observer's task and knowledge about a scene [STNE15]. Tasks like driving or playing a game has a strong influence on where we look and pay attention to. In a variety of experiments researchers have shown that recognition of only briefly presented faces is exceptionally accurate even if the participants have been forced to pay attention to other tasks [RMK07]. The same effect can be shown for a variety of other familiar objects [Gol13]. As a result, visual attention may be seen as a two-stage process beginning with a *pre-attentive stage* in which bottom-up features are quickly processed. The first stage is followed by a *focused attention stage* which integrates top-down attributes that are processed slower [Tre88].

Due to individual differences of how humans perceive a scene, and since emotion also can affect attention in a number of ways, a comprehensive model for the effects of attention on visual perception is still a topic of active research.

2.6 Summary

This section summarizes the findings from psychophysical literature which may play a role in perceptual models for computer graphics and gaze-contingent display algorithms.

- The horizontal field of view (FOV) of humans is about 160° for monocular vision and 200° for binocular vision. Vertically the FOV is about 135°.
- With visual acuity of $\approx$ 1 arc minute (20/20 Snellen) highest spatial frequency is perceived in the central foveal region. Resolving power decreases linearly with eccentricity. Very low spatial frequencies (< 0.1 cpd) cannot be perceived at all.
- Foveal vision is most sensitive to spatial detail and static contrast, whereas the periphery is most sensitive to motion.
- The temporal resolution of the HVS is limited by the critical flicker frequency (CFF) which varies with retinal illuminance, color and eccentricity. The CFF is about 40 Hz in the foveal region for normal photopic conditions and up to 70 Hz at 55° eccentricity.
- The dynamic range of simultaneously perceivable luminance covers 6.5 f-stops (without adaptation) and 46.5 f-stops when considering adaptation over time.
- Depth perception is based on disparity, vergence, accommodation (depth of field), parallax, contrast, texture and size. On 3D displays, the vergence-accommodation conflict may evoke visual discomfort and simulator sickness.
- The range in which the eye is able to accommodate ranges from approx. 8cm to ∞ and degrades with age.
- Paying attention to a region of interest at eccentricities >30° triggers a combination of head and eye movement so that the region is perceived at highest detail with the foveal region.
- The HVS processes motion by evaluating optical flow patterns in combination with signals from the vestibular system. Presenting an environment in a VR headset can cause motion sickness over time if the resulting flow patterns do not match the vestibular signals [MS92, PCC92, WHLP16].
- The eye is constantly in motion, performing fixations and saccades in order to sequentially scan the environment.
- Smooth pursuit eye motion keeps the projection of moving objects on the fovea. Tracking accuracy and perceived detail decreases with velocity of the tracked target.

- The distribution and connectivity of color-sensitive photoreceptors (cones) with higher-level retinal cells and the visual cortex is subject-dependent, non-linear and complex. For many different visual tasks the performance can be successfully approximated by linear functions.
- Differences in visual performance across the visual field can often be compensated by scaling the stimulus with projected eccentricity. The change in size can be described by an inverse linear function, called M-scaling or cortical magnification. The scaling parameters vary among visual functions.
- To equalize performance across the visual field, scaling along non-spatial stimulus dimensions, such as pattern contrast, is required along with size scaling [SRJ11].
- Levi's E_2 value is a useful yardstick for comparing the performance of different visual tasks [SRJ11].
- Peripheral vision can be improved for many tasks by learning. Perceptual learning is typically location-specific and affects basic visual functions, such as orientation discrimination, contrast sensitivity and some types of acuity [SRJ11].
- Recognition of scenes, objects and faces in peripheral vision does not generally follow predictions from cortical size-scaling and acuity measures [SRJ11]. This may be due to mid-level processes integrating local features into contours and other effects induced by attention. Additional scaling of non-spatial variables such as contrast may equalize performance [MR03].
- Directing gaze is a strong hint for selective attention. Attention may amplify or attenuate visibility of a stimulus.
- Low-level features (brightness, contrast, etc.) increase saliency resulting in a higher probability for directing attention to a certain region of interest. However, attention is influenced also by cognitive features such as scene knowledge and the observers's task.

Further Reading

This introduction only covers basic aspects of the Human Visual System. Excellent and detailed information is provided in Refs. [AKLA11, TFCRS11, Gol09, Gol13, CW11, WDW99].

Chapter 3

Related Work

Contents

The first part of this chapter summarizes strategies and available hardware in order to derive gaze information of a person (Sec. 3.1). This is relevant since all methods proposed in this work aim at gaze-contingency and, therefore, require knowledge about where the user is looking at. The second part of the chapter presents a variety of gaze-contingent rendering applications (Sec. 3.2). The section explains goals for each technique as well as respective state-of-the-art strategies.

3.1 Gaze Estimation

Gaze-contingent algorithms assume knowledge of the user's viewing conditions. Some information such as the fixation location on a display, the fixation duration and pupil dilation can be derived accurately by devices that *actively* track one or both eyes. Alternatively, *passive* gaze estimation strategies allow predicting gaze given the presented image or video without additional devices. However, passive gaze estimation can usually provide only a set of possible fixation locations or probabilities encoded in a saliency map. In addition, accuracy of gaze prediction depends on the complexity of the used model which, in return, has significant impact on the runtime of the tracking algorithm. Thus, gaze-contingent real-time applications usually require active gaze-tracking to be successful whereas passive strategies are more applicable to tasks without runtime limitations. Since algorithms in this dissertation cover real-time as well as offline applications, current solutions for both principles are briefly described in this section.

3.1.1 Active Gaze Tracking

Gaze-contingent rendering approaches require knowledge of where on the screen the user is looking at at any time. Depending on the application, gaze direction must be known with at least 1-2° accuracy and updated of 50 Hz while end-to-end latency may not exceed 60 milliseconds [LW07]. In most commercial eye tracking systems, small video cameras are mounted close to the eyes and record the user's eye balls. If the user is allowed to move the head, for example with a VR headset, the head must be tracked in addition.

Although actively used in many areas for research and product analysis, eye tracking devices are still emerging as a commodity technology [SMI16]. No current eye tracking product meets all requirements, such as accurate, robust gaze tracking capabilities with low latency, small physical dimension, low power consumption for mobile usage and low hardware costs [KHN16]. However, hardware components constantly become smaller and more efficient, indicating that next-generation eye-tracking technology will hopefully eliminate current limitations.

Measured Signals Depending on the type of eye tracker, the measured signal are the two rotation angles of one eye (monocular tracking) or both eyes (binocular tracking) relative to the head. Usually measured eye orientations are limited to rotations about the vertical (looking left or right) and horizontal (looking up or down) axes. Torsion, occurring when the eye ball is rotated about the optical

axis, has comparably low importance in most scenarios and is therefore mostly neglected. Some eye trackers can additionally measure pupil dilation [Res16]. Assuming a calibrated system, the measured eye rotations can be converted into the reference frame of the stimulus, e.g. the coordinate system of the display.

Gaze Measurement Techniques Eye-tracking algorithms have a long history. Due to the challenging task of estimating the gaze at high frame rates, the methods are often optimized for specialized setups which vary greatly in their design. A survey on eye tracking, including the employed eye models, can be found in [Duc02], [HJ10], [LU13] and [CY13], while methods to evaluate eye-tracking quality are presented in [HNM12] and [WMPH16]. Modern eye tracking systems can be categorized into three basic groups with regard to their functionality:

- *Magnetic search coils.* Systems based on magnetic search coils use contact lenses worn by the observer. Oscillating magnetic fields induce a changing current in the magnetic search coils. From the induced magnetic field the direction of the eyes is deduced.
- *Electrooculogram (EOG).* This gaze tracking approach uses electrodes placed below or above the eye. Starting from a neutral resting state of the eye, the measurable potential difference between the electrodes during eye motion is converted to angular values [MHO08].
- *Video-oculography (VOG).* This non-invasive technique is the most commonly used solution for tracking the eyes due to their comfort for the user when compared to both previous solutions. Spatial and temporal precision depend on multiple aspects of the tracking hardware. Stationary solutions for a fixed head position can achieve very high spatial solution ($<0.1^\circ$) and sampling rates of several hundreds Hertz to 1-2 kHz [Res16]. Mobile solutions are less precise due to noise from observer movement and environment lighting as well as due to constraints of the wearable camera hardware such as weight, physical size, temperature and cost. Recent mobile cameras achieve performance levels sufficient for many gaze-contingent applications.

In the following, video-based techniques are described in greater detail. A taxonomy of the different tracking algorithms is given in [TA13].

Tracking the Pupil-Corneal Reflection Most popular, non-intrusive approaches make use of a distinct, bright pupil-corneal reflection which can be detected robustly (Fig. 3.1). An infrared light source and camera are directed towards the user's eye, and the captured images are used to determine gaze direction by measuring the position the bright corneal reflection in relation to the dark pupil [Duc07]. This Purkinje reflection of the cornea is the bright "glint" appearing in the recorded infrared image (Fig. 3.1, right). The bright corneal glint remains stable due to the spherical shape of the eye ball, whereas the pupil follows the gaze direction [MM13]. Under optimal conditions, such as precise calibration and a clearly visible glint, gaze tracking exploiting the corneal reflection is able to achieve optimal spatial resolution (mean error $< 0.1^\circ$) [HJ10]. It is widely used for perceptual studies.

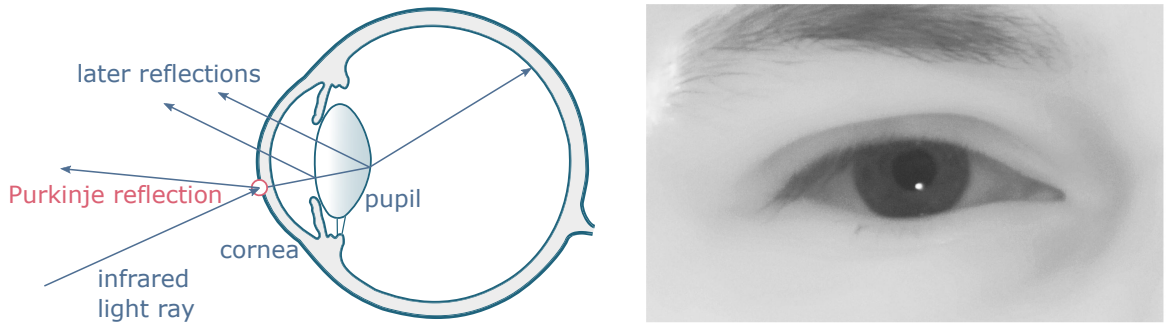

Fig. 3.1 Purkinje reflection. (left) The incoming infrared light is reflected off different parts of the eye. The first and brightest reflection (Purkinje reflection, "glint") from the outer surface of the corena is commonly used for tracking and must be prominently visible from the tracking camera (right).

However, usage of the corneal reflection "in the wild" is difficult due to serious robustness issues. If the corneal reflection is ambiguous due to additional optics, such as if the user wears contact lenses, or if the reflection is physically occluded, e.g. by the frame or the tracking device, tracking is less precise or even completely lost. With reflection-based systems, the main challenge is to deal with ambiguous glints and reflections [KKS09, DBBS06], blinks [CE14], or noise [LWP05], especially for wearable, near-eye devices [CE14, BJ11].

Feature-based Eye Tracking Feature-based methods do not extract a single peculiarity from the image but use the complete input image to estimate the gaze. The basic feature-based gaze tracking pipeline is as follows: First, the pupil contour is extracted from the camera frame. From the contour the pupil center is derived and converted into screen coordinates using user-specific calibration data. This approach is beneficial when the Purkinje reflection cannot be robustly detected, which may be the case in a VR headset with complex optics. LED illumination in front of the lenses would result in visible reflections from the lenses themselves, while a placement behind the lenses and close to the eye has the drawback that glints will not be visible for the entire wide FOV. Similar to reflection-based approaches, also feature-based eye tracking methods illuminate the eye using infrared light to enable eye tracking without impairing the user's viewing comfort.

An essential step for feature-based eye tracking is to robustly detect the pupil. The pupil forms the darkest part of the eye if illuminated from an off-axis view, and the brightest part if illuminated from a near-camera-axis view [HNM12]. Consequently, it is often well-separable from the surrounding iris. Most techniques rely on edge or contour detection of the pupil, followed by ellipse fitting.

Several approaches build upon this idea, e.g., in form of multi-layer networks [BP94], Gaussian processes [WBC06], or manifold learning [TKA02]. While being flexible and requiring only a calibration step, these techniques are often computationally costly and less applicable for scenarios where high tracking rates are required. Recent research enables faster (10-15 Hz), calibration-free eye tracking on mobile devices with moderate accuracy [KKK+16]. Krafka et al. make use of a

convolutional neural network which is trained using large image databases of eye regions. Other systems include a pair of calibrated and synchronized cameras to triangulate the position of the eye in space. Gaze direction is derived from features of the pupil and eye visible in both cameras. Due to the calibration of the cameras, a two-camera system may work without user calibration, but at increased hardware cost and size limitations.

In Chapter 6 a novel approach is presented that combines image features with an underlying simulation model of the proposed HMD. The pupil is detected in the recorded views, and the gaze direction is derived using a physical eye model.

Calibration Video-based eye tracking methods require an individual calibration phase of the hardware for each user. For this task, traditionally, a defined grid of 3x3 or 4x4 markers is sequentially presented to the user. The tracked eye is recorded during fixating each marker, resulting in a recorded mapping from eye direction to screen position for the marker set. After the calibration procedure, in-between eye rotations are interpolated so that a screen position can be derived for every possible eye direction. A comprehensive guide to eye tracking calibration methods has been proposed by Holmqvist and co-authors [HNA+11].

Importantly, the calibration is only valid if the transformation between the head and the eye tracker does not change during tracking. Otherwise, the derived gaze direction drifts away from the correct position on screen. Stationary eye trackers therefore usually use a chin rest to keep the observer's head in place for the duration of the experiment. Similarly, for head-mounted displays with eye-tracking capabilities, any sliding movement of the HMD due to head motion during data acquisition will result in erroneous data.

Potential drift and the time-consuming calibration process hamper deployment of gaze-contingent applications. Just recently, researchers have begun to loosen calibration constraints to enable eye tracking also for more general scenarios. Cazzato et al. derive pupil information and head pose at the same time to enable gaze tracking while both the eye tracking camera and the tracked person can move independently [CLD14]. The presented eye tracker in Chapter 6 simplifies marker-based calibration to a single marker.

3.1.2 Passive Gaze Tracking and Gaze Prediction

When we look at a scene our scanning eye movements are not random. When observing a given image, different people tend to look at similar points, in average [CW11]. Visual perception research has discovered gaze patterns that are common for healthy adult humans, although differences exist between cultural environments [CBN05] and gender [VCD09, SI10]. Humans are similarly attracted by faces and objects that are located in the line of sight of such faces [Gol09]. Painters intuitively exploit the appeal of faces. Analyzing scan paths using active eye tracking has revealed more similarities [CW11]. Many of the common gaze properties can be explained in an evolutionary content. Being attracted by an item that features a certain attribute that is distinct in a group of items, e.g. the color of a fruit in

an environment of a different color, increases efficiency when searching for food. Recognizing and instantaneously looking at a moving predator lurking in the covert ensures survival.

Methods for passive gaze tracking and gaze prediction aim at modeling these findings and try to estimate for a given image or video where people will look. Some algorithms try to also derive the order of fixations to scan the presented visual stimuli [BI13]. The probability that a certain image region is actually observed is usually encoded in a *saliency map*. Inspired by feature integration theory [Tre88], the saliency map can be thought of as a summary of the conspicuities of all visual stimuli.

In the past two decades the variety and quality of gaze prediction methods has increased dramatically. A recent survey on visual attention modeling has been proposed by Borji et al. [BI13]. Most models used in gaze prediction can be categorized into two groups, *bottom-up* models and *top-down* models. In correspondence to the psychological literature, these models are either driven by basic visual stimuli of the HVS, such as contrast, edges or boundaries (*bottom-up*), or they are driven by the task and intention of the subject understanding the scene (*top-down*).

Bottom-up Prediction

From the bottom-up perspective, values in the saliency map are normalized center-surround differences which are computed for individual stimulus features and added linearly. The model from Itti & Koch is one of the first and most-cited computational methods following this idea and applicable to images and videos [IKN98]. Being biologically inspired, the model measures local center-surround contrast on different scales simulating the receptive fields of Ganglion cells in the retina and neurons in the visual cortex. Images are first resampled into a multi-resolution representation. Each image is then separated into an intensity channel and two color-opponent channels for red-green and blue-yellow contrast, respectively. The intensity image is then filtered to obtain gradient orientation maps at four orientations. For each feature, such as color, intensity and orientation, the difference of Gaussians is evaluated on different scales of an image pyramid. Respectively, information across two resolution levels is integrated into so-called feature maps. Finally, these feature maps are accumulated over all resolutions and integrated into one common saliency map. The Itti & Koch model can be quickly evaluated so that modern implementations allow estimating saliency in real-time.

One efficient implementation has been proposed by Longhurst et al. which runs on the GPU but relies on computer-generated scenes. The approach extends the Itti model by evaluating also depth, motion and habituation components [LDC06]. The additional model components are motivated by the observation that objects closer to the observer are more salient. Habituation refers to the effect that objects become familiar over time.

Mannan et al. observe increased saliency for regions of high edge density [MRW96] whereas Parkurst & Niebur noticed that saliency increases with luminance contrast [PN03]. Harel et al. proposed a graph-based framework that adapts the Itti & Koch model for combining low-level saliency features [HKP07]. In contrast to previous top-down approaches, the authors observed that objects in

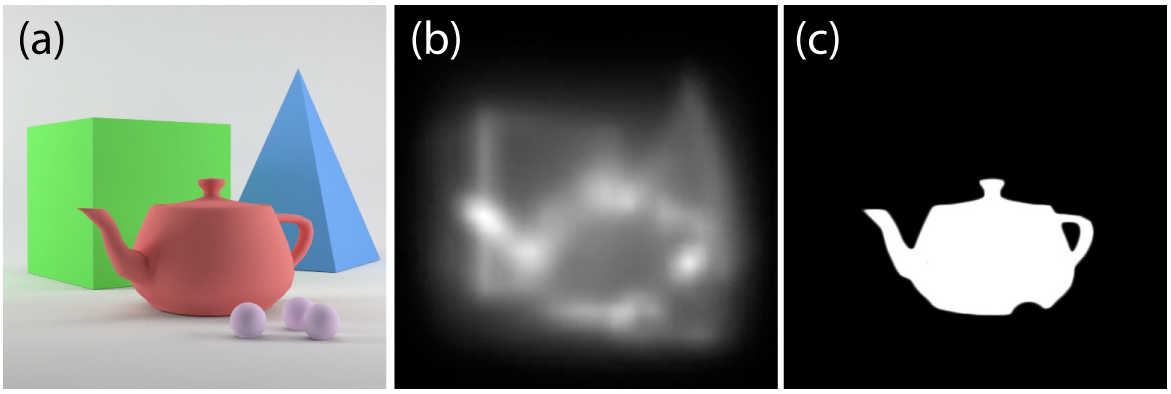

Fig. 3.2 Bottom-up and Top-down Saliency. Given image (a) task bottom-up saliency using [HKP07] may predict fixations in a free-viewing task (b). Top-down prediction in a visual search task for the teapot may result in (c).

the central part of the visual field seem to be more salient. Therefore, the model includes a center bias, yielding better results than previous approaches.

Jansen et al. additionally model the influence of binocular disparity on saliency, related to other visual cues resulting in perceived depth differences [JOK09]. The original model of Itti et al. gives high value to edges of all frequencies. However, human vision is less sensitive to low-frequency edges which is included into the saliency predictor of Murray et al. [MVOP11]. This approach uses an inverse wavelet transform over center-surround output which elegantly includes scale into the computation process. The model is initially trained by using eye fixation data and color appearance measurements which reduces the number of free parameters and gives better results when compared to many other bottom-up predictors.

Alternatively, visual saliency may directly be learned from large amounts of eye tracking data [ZK13].

Top-down Prediction

Top-down methods model scene understanding due to the observation that humans are biased to object features during performing a particular tasks [NI07]. Hence, top-down attention models commonly introduce a visual feature bias with respect to known objects in the scene [MTT04]. The saliency methods briefly summarized in this section derive scene knowledge from figure-background segmentation [FWMG15], face detection [VJ04], person detection [FMR08], object detection [CHEK08] or manually defined task-specific location bias [CCW03] (Fig. 3.2c). Top-down gaze prediction is usually used in combination with bottom-up approaches in order to derive the overall salience of a pixel resulting in higher prediction accuracy for task-based scenarios.

Itti and Koch propose a simple predictor which uses a weighted sum of the saliency of all feature values [IK01]. The weights represent multiplicative gains for task-related features.

Gaborski et al. try to learn the relationship between task and image. A neural network is fed with color opponent images and task information [GVC03]. Similarly, Sundstedt et al. introduce an importance map for task-relevant objects which is combined with a bottom-up saliency computation step [SDL+05]. The authors used this model for selective rendering and performed a task-based experiment. The task of the participants was to find and count all fire safety items in a given scene. Interestingly, the group of tested people were not able to distinguish a high-fidelity rendering from selective rendering results based on top-down features.

Walter & Koch introduce the proto-object map, derived from edge intensities of detected objects, in order to model the top-down component [WK06]. Navalpakkam et al. explicitly take the effect of distracting visual features into account. Therefore, the authors use knowledge about the scene object to model the optimal top-down bias by modulating feature gains with respect visual search tasks [NI07]. For example, if the task is to search for an upright green bottle, the method increases saliency for the known object properties such as orientation and color in the specific case. Bottom-up information about the salient object and distractor objects in the background is provided and used for learning the optimal distribution of feature gains.

Cerf et al. examine gaze behavior for natural scenes [CHEK08]. First, a face detector is applied to the image. The hybrid approach increases saliency values of a bottom-up saliency map for the detected face locations, and yields good results.

Judd et al. follow the idea that saliency depends on task- and scene-dependent cues as well as on bottom-up saliency cues. Therefore, the authors suggest to learn where people look directly from eye tracking data. The authors use a data base of over a thousand natural images in an eye tracking study with a free-viewing task to generate ground truth saliency data. From the images different features are automatically collected including low-level (intensity, orientation), mid-level (vanishing point, horizon line), and high-level features (face and object detection). A linear support vector machine is then used for training the saliency model. In a proposed saliency benchmark the technique outperforms previous approaches [JEDT09].

Han et al. compute top-down saliency in a probabilistic way from the given image. From every patch of the image an "objectness" likelihood is derived by using a trained model. For training, a large number of eye-fixation patches from an eye-tracking dataset are used [HHQ+13]. Frintrop et al. combine the Itti & Koch with an object proposal generation framework [FWMG15]. Each detected object segment is flood-filled with the respective local maximum from bottom-up saliency resulting in a segment-based saliency map. With their work the authors show that in combination with a top-down component, the bottom-up model by Itti & Koch is still competitive to other computationally more complex methods.

Recently, deep convolutional networks trained on large image data sets have shown impressive improvements for fixation prediction [VDC14, KTB14, KAB15]. In comparison to dedicated feature detectors, trained networks are able to better model the influence of high-level features (faces, text) and abstract features like popout. Kümmerer et al. reuse existing neural networks to decrease the computational effort to create a network for saliency prediction [KTB14]. With the "Deep Fix"

network Kruthiventi et al. introduced location-biased convolutional filters which enables the deep network to learn location dependent patterns of fixations, such as the center bias observed by Judd et al. [KAB15].

Fixation Location Prediction Quality

Attention models for passive gaze prediction – no matter which strategy the algorithm follows – do not provide exact solutions. In terms of accuracy, fixations from saliency maps are not comparable to active gaze tracking. However, knowing the approximate gaze location may provide a non-invasive solution that is sufficient for some applications. The prediction accuracy for bottom-up saliency is typically evaluated by a free-viewing task in which participants watch photographies and videos for the very first time [JDT12]. However, there is some controversy about the role of bottom-up versus top-down mechanisms in the context of gaze prediction [JDT12, VDC14, KWB14, BJD+15].

Perceptual experiments have observed how well bottom-up and top-down models can predict fixation locations [JDT12]. Free-viewing experiments assume controlled conditions to be comparable which is difficult to achieve since participants may be biased by cognitive load when performing the tests. The influence of the task on bottom-up saliency prediction has been the focus of a variety of studies. By performing eye-tracking experiments Cater et al. have shown that many low-level features become irrelevant if they do not contribute to a certain task [CCW03]. The authors explain their findings by the inattentional blindness effect [MR98] (cf. Sec. 2.5). As a result, salient regions can fail to capture attention if they conflict with the viewer's goals.

Sundstedt et al. observed that fixations and saccades are completely different for task-based tests in comparison to free-viewing experiments [SC06]. The authors tested the quality of bottom-up saliency (Itti & Koch [IKN98]) vs. task-based saliency (Navalpakkam & Itti [NI02]) used for region-of-interest prediction in selective rendering. In the experiments 64 participants were either performing a task or were freely viewing an animation. Confirming the results of previous work, eye movement should correlate with the task map (top-down) when performing a task, or with the bottom-up saliency map when freely exploring the scene. Indeed, with 80% of the fixations being located in a 4° area of the task-map maxima, the top-down predictor achieved good accuracy. Without a task, participants were guided more by low-level features and paid more attention to salient parts in the periphery. Sundstedt et al. conclude that top-down saliency is a much better predictor if the task of the user is known.

Einhäuser et al. confirmed that for task-based scenarios, top-down mechanisms override low-level features so that bottom-up saliency maps become irrelevant [ERK08]. The authors show that for free-viewing conditions, bottom-up saliency methods robustly predict the first two or three fixation locations for natural scenes. Bottom-up saliency models fail to correctly predict gaze for succeeding locations as higher-level processes of image interpretation set in and, top-down methods achieve more accurate results.

There exist stimuli that capture attention regardless of performed task, such as the sudden appearance of an object which generally captures attention [KSR+03]. The involuntary transient shift

from the previously attended object to the appearing object occurs even when the part of the scene is uninformative or may impair task performance [CS02].

In several other studies the overall results have been confirmed also for natural scenes [GVC03, SU07, STNE15]. Low-level saliency maps correlate with fixation locations for free-viewing tasks but achieve low accuracy if the participant performs a task. In this case top-down features dominate low-level features .

In a large-scale benchmark, Judd et al. measured average fixation locations of 39 participants for a database of 300 natural images and tested ten different saliency detectors. The benchmark shows that the predictors of Harel et al. and Judd et al. achieve good prediction accuracy for a larger number of images. However, the benchmark tests indicate that there is no single method equally suitable for all types of scenes and situations [JDT12]. Accuracy is significantly increased when combining the results with an additional face detection step. The authors also tested gaze similarity with a varying number of ground-truth data sets. As a result, gaze data from two observers already gives more accurate results than the best-tested bottom-up gaze predictors. A gaze data set of 10 participants already contains a large part of the whole ground truth data. Later benchmarks using different metrics are provided by Vig et al. , Kümmerer et al. and Bylinskii et al. and show that for free-viewing tasks saliency prediction based on convolutional networks learned from gaze-labeled natural images often outperforms traditional "hand-crafted" saliency predictors [VDC14, KWB14, BJD$^+$15].

Scan Path Prediction

Saliency prediction seldomly results in a single, distinct salient region. To estimate the sequence of fixation locations of an observer is therefore a much harder problem and has largely been ignored for a long time in saliency research [NSEH10]. Scene viewing models have primarily been designed to predict *potential* fixation locations. One naïve approach for a given input video is to apply saliency estimation for each video frame separately and to search for the highest saliency value per frame (*winner-take-all* principle). The concatenation of maxima gives the scan path over time. However, this strategy performs poorly in many situations. First, temporal stability is not modeled which ends up in jumpy, unnatural gaze paths. Second, psychophysical experiments have revealed the *inhibition of return* which also affects the resulting scan path. The inhibition of return explains why the eye moves to all potentially interesting locations of the scene instead of fixating only on the most salient region. Chua et al. have shown that cultural differences may result in significantly different scan paths [CBN05]. In the tested photographs with a focal object on a complex background westerners fixated more on focal objects whereas East Asians attended more to contextual information.

Approaches for scan path prediction have been developed for reading, photography viewing and watching video. Commonly, the techniques mimick human scene viewing in terms of discrete temporal phases with fixations when the point of regard is relatively still and saccades when the eye switches gaze from one location to another. Fixation durations have been first modeled for reading tasks. For scene viewing, selection prediction of the next fixation location turned out to

be harder [NSEH10]. Henderson et al. confirm that scan paths generated by bottom-up saliency maps do not well correlate to ground truth [HBCM07]. Recorded eye-tracking data in combination with saliency and attention maps achieves much better results for predicting the scan path for scene viewing and reading tasks [DCK13]. Nuthmann et al. significantly improved computational modeling of fixation durations for scene viewing [NSEH10, NH12]. The authors present an algorithm to automatically compute scan paths including gaze locations, fixation durations, and inhibition of return. Fixation durations are modeled as continuous-time random walks. The achieved results correlate to ground truth better than chance.

Dorr et al. present approaches for extracting the scan path from a given video using machine learning on gaze data in combination with a perceptually inspired color space [BDK$^+$06, DMGB10, DVB12]. The comparison with ground truth data from eye tracking show that, depending on the type of video, a high variability in scan paths can exist between subjects. The lowest variability has been computed for professional-grade movies, due to gaze-guidance strategies of the movie director. Many professional videos are being viewed in quite a predictable fashion and have a single and well-determinable salient region attracting the attention of the viewer while remaining regions mainly remain unattended [BMS02]. With respect to previous findings and studies on a large video dataset, Dorr et al. hypothesize that bottom-up saliency values can aid in determining a set of potential saccade targets. The actual saccade target is selected based on the history of previous saccades and on other mechanisms (inhibition of return, task-specific bias, etc.) [DVB12].

Where, and in what order fixations take place remains a challenging problem, especially if no ground truth gaze data is available. Robust, accurate and temporally stable scan path prediction remains a topic of ongoing research [VDMB12, VDC14, HLSR14, NE15]. In the future, techniques that account for both the *where* and *when* decisions will increase gaze prediction accuracy. In addition, in comparison to work on image saliency there exists significantly less research on video saliency, although motion and moving objects are known to be strong attractors of visual attention. Little research has so far been invested into saliency prediction in the outer peripheral field of view, i.e. the kind of visual stimuli that are able to elicit long-range saccades.

Further Reading

Detailed information on computational models of visual processing and eye tracking is provided in Refs. [LM91, Ray92, Und98, VG07, Duc07, Gol09, CW11, HEKR14].

3.2 Gaze-contingent Applications

The notion of gaze-contingent display devices dates back at least two decades. Due to the vast amount of related work in the areas of perceptual graphics and gaze-contingent rendering, including diverse topics such as gaze interaction, perceptual studies, attentive user interfaces and teleconferencing, gaze animation, selective rendering for ray-tracing, view-dependent geometric level-of-detail, visual equivalence for shading, video transmission bandwidth reduction, display resolution enhancement, gaze guidance, gaze-contingent tone mapping and depth of field, this section focuses on the most closely related work. In the following, some of the key contributions of each topic are briefly presented. Excellent review articles on gaze-contingent techniques and applications include those of Reingold [RLMS03], O'Sullivan [OHM+04], Duchowski [DCM04, DÇ07], Dietrich [DGY07], Bartz et al. [BCFW08] and Masia et al. [MWDG13].

3.2.1 Perceptual Studies

Gaze-contingent displays balance the amount of rendered visual detail against the perceivable visual detail over the FOV. Perceptual studies allow creation and validation of perceptual models required to drive gaze-contingent algorithms [RLMS03].

Parkhurst and Niebur investigate how gaze-contingent level-of-detail rendering affects our ability to detect and localize objects in visual search tasks [PLN00, PN04]. The authors propose a gaze-contingent image rendering approach which blends two different resolutions of the image. Only the foveal region is displayed at high resolution. In a perceptual study the authors evaluate performance and fixation durations during a visual search task. Their experiments demonstrate that object detail significantly influences the speed with which we are able to perform different tasks. By adjusting level-of-detail parameters conservatively, task performance can be normalized. With a central high-resolution area of 5°, task performance and fixation duration is close to the normal behavior at full resolution over the entire FOV [GPN06].

McConkie and Loschky have investigated post-saccadic perception when sensitivity is greatly reduced by saccadic suppression [LM00]. The study shows that displayed information cannot be perceived until 6ms after a saccade. Perona et al. studied detection of animated objects in briefly presented scenes. The authors found out that an animated object in a static environment can be detected in less than 27ms [FFIKP07].

A real-time simulator of glaucoma and other ophthalmic degradations of the FOV has been presented by Rayner et al. [RB79]. A related system by Perry and Geisler accepts conventional video footage as input and filters it with a pre-defined low-pass kernel centered on the current gaze direction at 60 frames per second (fps) [PG02]. Additional studies on glaucoma and macular degeneration confirmed the suitability of gaze-contingent displays for simulating visual defects as long as latency is low [RB79, FR99, MG09, VAS08].

Dedicated measurements to determine acceptable latency for gaze-contingent displays have been conducted in several studies [LW07, SRIR07, SW14, RJG+14]. The measured *end-to-end latency* comprises the full gaze capture and rendering pipeline, starting with capturing the frame for eye-tracking and ending with the reception of the photons emitted by the display and received by the photoreceptors in the retina. The gaze-contingent display system of Santini et al. which renders at 200 Hz achieves an end-to-end latency of only 10ms with dedicated hardware [SRIR07]. Loschky et al. observed that the display has to be refreshed after 5 ms to 60 ms after a saccade for an image update to go undetected. The acceptable delay depends on the task of the application and the stimulus size in terms of induced peripheral degradation. Beyond that time delay, detection likelihood rises quickly [LM00, LW07].

In recent work, Mauderer et al. created a model for simultaneous contrast perception [MFN16]. The approach modulates the color of the scene in the periphery according to the gaze direction which results in more saturated color perception. The authors plan to use the effect to create a new form of high dynamic range images with increased perceivable gamut size [MFN16].

Recording and analyzing gaze data for modeling and training saliency algorithms is a non-trivial task due to the spatially and temporally high-frequency nature of gaze data. Blaschek et al. provide an excellent review of the latest and methods for visualizing gaze tracking data for perceptual studies and gaze-based applications [BKR+14]. A testbed for gaze-contingent visualization techniques with respect to contrast sensitivity, color degradation and depth of field has recently been provided by Bektas and colleagues [BCKD15].

3.2.2 Attentive User Interfaces

Using gaze as an interaction metaphor is intuitive for search tasks but also turns out to be ambiguous and error-prone when being used for selecting or triggering commands [Jac91]. Special graphical user interfaces reduce ambiguities for gaze writing tasks but have not been able to reach interaction bandwidths that are competitive to established input devices such as the keyboard [WRSD08, PT08, MHL13]. Under normal conditions the eye is used to gain information about the environment but not to trigger commands. However, different studies have shown the gain in task performance, if gaze is combined with other modalities such as touch or head gestures [SD12, MHP12].

Another topic of attentive user interfaces is immersive video-conferencing. Vertegaal et al. developed the eyeCONTACT video-conferencing system which uses a camera array of three cameras to provide a video stream showing the user parallax-free and with a central gaze direction to increase the feeling of presence for the observer [VWSC03]. In another system, Chen and colleagues observe that eye contact between different people is assumed *a priori* as long as the gaze direction approximates the viewed person. The system by Chen et al. proposes design parameters for teleconferencing systems to enhance presence by increasing the plausibility of interpersonal eye contact [Che02]. Similar systems have been proposed by Fuchs et al. using a “sea of cameras” to enable depth and photometric

computations. Their system renders multiple users in a single immersive virtual environment in real-time [FBA+94].

Different systems also make use of real-time eye tracking to enable glasses-free autostereoscopic displays [AHEF02, Lee09, BHSS15].

3.2.3 Avatar Animation

Various studies have shown that animating the eyes and gaze of virtual avatars increases induced immersion in VR and gaming applications [GSV+03, MR06, SWM+08]. In VR headsets, real-time gaze animation can be achieved by video-based eye tracking [SMI16] or by electrooculography (EOG) [MHO08]. EOG-based HMDs have the advantage that they do not reduce the available FOV but they provide less accurate tracking results.

High-fidelity facial animation is usually captured with complex hardware such as a light stage or a head-mounted facial capture system [AFB+13, JFY+11]. For digital doubles the captured surface is mostly manually rigged and rendered offline [AFB+13]. A novel approach by Bermano and colleagues enables a temporally coherent reconstruction of the detailed eyelid geometry and skin wrinkles [BBK+15].

Just recently, optimized algorithms using RGB(-D) camera data as input allow capturing a face and controlling a completely different avatar by performance-based facial animation in real-time [WBLP11, CWLZ13], or even transfering the captured facial movements to an actor in a video [TZS+16].

With current VR headsets, the face is largely covered by the HMD so that traditional face capture methods are not applicable. Instead, Li et al. use strain sensors attached to the HMD and a depth sensor to capture the mouth region in order to reconstruct facial expressions in real time [LTO+15].

Epic Games built a pipeline to capture the full body of an actor, including facial expressions and gaze, to transfer the performance onto a virtual human including realistic rendering, all in real-time [Gam16]. However, the complexity of the required hardware and capture pipeline does not allow to use these techniques in other fields beyond professional movie or game production.

Another novel direction of research for primarily gaze-based avatar animation is autonomous perception of virtual humans. Neog et al. simulate visual perception by generating saliency maps from the viewpoint of the avatar [NCRP16]. The gaze direction of the avatar is controlled by the most salient region. In addition, high-level attributes control the expression of the eye region of the avatar, e.g. anger or surprise, as has been shown for “Digital Emily” watching a hockey game [NCRP16].

3.2.4 Selective Rendering

Display algorithms for selective rendering save rendering cost in perceptually less important regions of the image. As early as 1990, Levoy and Whittaker proposed a gaze-contingent approach to render volume data sets according to view direction [LW90]. Motivated by limited memory and computational resources, a ray tracer was described whose local ray density varies depending on the angle between volume region and gaze direction, while an eye tracker continuously measures gaze direction in real-time.

Selective rendering is commonly used in ray tracing to steer the number of samples per pixels or recursion depth [FPSG96, SDL$^+$05, LDC06, HCS10]. A common goal of selective rendering is to obtain an image that is perceptually indistinguishable from a fully converged but computationally expensive rendering solution. Myszkowski and colleagues use the perceptually motivated *Visual Difference Predictor* as an image metric to selectively stop rendering of a Monte Carlo path tracer [Mys98, HMYS01]. In another approach, Farrugia et al. make use of a perceptually-inspired metric based on eye adaptation for a progressive rendering method to stop early-exit global illumination computation [FP04].

Modern selective rendering methods target high-fidelity, offline rendering. The perceptual importance of the final image is usually approximated by saliency extracted from previews rendered at lower quality [YPG01, CCW03, SDL$^+$05, CDdS06, HCS10, GDS14, Har16]. The initial image estimate requires at least one sample per pixel [YPG01, CCW03]. To decrease creation times, Longhurst et al. present a faster approach by computing the preview frame via rasterization. The preview is then used to extract saliency including different visual cues such as edges, intensity, motion, depth, color contrast and scene habituation [LDC06].

For selective rendering, the generated saliency map is used to steer the number of samples distributed across each pixel of the image. Saliency is created by established strategies that are based on low-level features and on object properties, which are known for rendered scenes. Cater et al. present a purely task-based approach for selective rendering [CCW03]. In this approach objects are annotated to define which objects are relevant for a certain task. The object importance is assigned to each object resulting in a rendered task map which is then used for steering rendering quality. The authors performed a perceptual study using eye tracking to validate their approach. Chalmers et al. investigate several ideas, such as importance-based sampling for on-screen distractors, e.g., sound-emitting objects, or sorting of effects, to compute the visually most important paths first and postponing less important reflections or global illumination [CDdS06]. Hasic and colleagues show the importance of visual tasks and motion for selective rendering because both attract the viewer's attention [HCS10]. In addition, Yee et al. take motion sensitivity of the HVS into account and increase saliency accordingly for moving objects [YPG01].

Perceptual models for selective rendering are capable of simulating several properties of the HVS but are often too costly for real-time rendering [FPSG96]. In addition, highest quality image quality is currently achieved by offline ray-tracing only. However, hardware-accelerated ray-tracing

systems can achieve interactive rates, depending on the scene complexity [PKC15]. Recently, Fujita et al. presented a real-time ray tracing system for VR headsets. The authors exploit the perceived blur in the periphery due to the distortion of the HMD optics to reduce the number of traced samples in the non-central parts of the screen to increase rendering speed [FH14].

3.2.5 Gaze-contingent Level-of-Detail

Researchers use perceptual models also to reduce the number of polygons in areas of lower acuity, making view-dependent geometric level of detail (LOD) another example for gaze-contingent rendering [OYT96, Hop98, HSH10].

In their 1996 paper, Ohshima and collaborators employ gaze-aware LOD rendering in order to interact with multiple objects in a virtual environment [OYT96]. Besides angular distance from gaze direction, the authors take additional perceptual clues from kinetic and binocular vision into account to adapt the rendered level of detail to what can and cannot be perceived. In contrast, Luebke et al. simplify 3D geometry meshes directly in accordance with gaze [LH01]. To remain visually imperceptible, the degree of mesh simplification is controlled by a perceptual model and eye tracking data. Along similar lines, Murphy and Duchowski propose a non-isotropic LOD rendering approach using eye tracking for geometry meshes based on a user study-derived 3D spatial degradation function [MD01].

Gaze-contingent LOD has proved especially beneficial for visualizing highly tessellated terrains to reduce geometric detail not visible to the user [SLL+14, Red01]. The approach of Reddy et al. drives the LOD for terrain meshes based on an image frequency analysis [Red01]. This analysis is derived from perceptual observations and is applicable to images rendered from different viewpoints and with different LODs. Instead of discrete LODs, the approach of Williams et al. performs adaptive mesh simplification [WLC+03]. This method evaluates local simplification operations using a contrast sensitivity model based on visible silhouettes, highlights and pre-processed static textures.

Parkhurst and Niebur measure visual search times to find the optimal peripheral geometric LOD reduction. The derived model includes an eccentricity-based acuity fall-off and pixel velocity during during navigation [PN04]. The results are confirmed in a study by Duchowski et al. [DBS+09]. In contrast to the work by Parkust et al. , the latter study varies image quality with respect to perceivable colors across the visual field. The results imply that color detail cannot be reduced as readily as geometric or pixel detail [DBS+09]

Sundstedt et al. selectively render task-relevant objects and other salient features in high quality and reduce rendering quality for the remaining parts. In a visual search task the tested people were not able to distinguish high-fidelity rendering from selective rendering results. The experiments impressively demonstrate the suitability of perceptual rendering if selective attention can be predicted [SDL+05, SC06].

The mentioned geometric techniques drastically reduce the workload for geometry processing. At the time of publication the techniques resulted in speed-up factors of one or two orders of magnitude

due to the limited rasterization performance of the available graphics hardware. However, in modern rasterization shading time has become the major bottleneck [HGF14]. Hence, besides geometry reduction, adaptation of the shading quality is another important point of consideration for gaze-contingent rendering.

3.2.6 Gaze-contingent Shading

In current pipelines for real-time rendering, shading is often more expensive than the geometry pass [VST+14, HGF14]. Multi-rate and multi-resolution shading are novel strategies in real-time applications that enable adaptation of shading complexity to scene content.

Early works in gaze-contingent multi-resolution rendering primarily analyzed the detectability and influence of the quality degradation on visual performance [PP99, PLN00, NNB+04, DBMB06]. However, these approaches have not brought a boost in rendering performance.

Foveated 3D graphics (F3D) simulates acuity fall-off by rendering three nested layers of increasing angular diameter and decreasing resolution around gaze direction [GFD+12]. These are then blended into the final image. F3D achieves impressive shading cost reductions but also introduces overhead by repeating rasterization for each nested layer. Nvidia recently proposed a multi-resolution shading approach by drawing different resolutions within a single pass on their newest GPU hardware [NCR15]. The approach exploits the inevitable image distortion in HMD setups and draws the image at different resolutions, subdivided into a fixed 3×3 grid. The reduced resolution saves between 20% to 50% of pixel shading cost and can perform even better if combined with F3D. However, single-pass, multi-resolution rendering requires special multi-projection GPU functionality.

The idea of adaptive *Multi-rate shading* (MRS) is to distribute more shading samples near object silhouettes, shadow edges, and regions of specular highlights. In blurred regions, induced by motion blur or depth-of-field, shading samples are distributed more sparsely [HGF14]. The approach achieves impressive savings in terms of shaded fragments (50% to 80%) without reducing perceived render quality. Efficient implementation of this approach requires an extension of the graphics pipeline and is currently not feasible on commodity graphics hardware.

Coarse pixel shading (CPS) enables different shading resolution levels by executing shaders at three varying rates: per pixel group, per pixel, and per sample [VST+14]. Results within a software renderer show shading savings comparable to MRS and applicability to foveated rendering. Vaidyanathan et al. tested their approach for foveated rendering using a simplified acuity model. Assuming static gaze and a constant radial acuity function, shading is computed at full-resolution in the foveal region and at a lower rate outside towards the periphery.

In a practical implementation of foveated rendering using the Source Engine™, the Valve Corporation recently presented a 10-15 % boost in rendering performance [Vla16].

Swaffort et al. provide an image metric for perceptual foveated rendering quality [SCM15, SIGK+16]. The metric extends the HDR-VDP2 predictor [MKRH11] by measuring peripheral vision degradation. In the novel metric, contrast sensitivity function decreases with visual eccentricity

based on the tuned cortical magnification factor (CMF, Chapter 2.3). The study proposes different sets of parameters for multi-resolution (foveated) rendering and multiple established shading effects. Optimized parameters are derived from a perceptual study and allow tuning the CMF for the foveated image quality metric.

3.2.7 Accommodation Simulation

In an environment with objects at different depth accommodation of the eye adjusts the focal distance to the point of regard (POR). Objects projected onto the area of the fovea are perceived clearly. Other objects appear increasingly blurry with depth from the focal distance [AKLA11]. In a typical gaze-contingent display, the user perceives a sharp image everywhere because the eye accommodates to the distance of the screen. However, a rendered 3D scene or a 3D movie usually contains objects at different virtual distances. Perception of the depth at the POR and at the same time accommodation to the (probably different) screen distance results in the accommodation-vergence conflict (AVC, compare Fig 2.17). This discrepancy increases visual discomfort, fatigue, and can even cause motion sickness [HGAB08]. Studies have shown that gaze-contingently rendered defocus, also known as depth-of-field (DoF), may reduce some negative side-effects of AVC [VAF16]. However, due to the increase in blurriness of the resulting image, some subjects also dislike the feature and rate image quality as being lower compared to an image without gaze-contingent DOF [DHG+14, VAF16]. Alternative approaches are discussed in the survey paper by Kramida et al. [Kra16] including hardware supporting natural accommodation such as auto-refracting lenses [LHH+09, LHC10] and multi-focal displays [NAB+15, HLW15], or tuning the disparity of all scene content so that the virtual distance of the POR is shifted into the focal distance of the screen [TDM+14, KDM+16].

Excellent surveys on DOF rendering methods are provided by Demers et al. [Dem04], Barsky et al. [BK08] and McIntosh et al. [MRD12]. Studies on gaze-contingent DOF rendering are available from Hillaire et al. [HLCC08], Mantiuk et al. [MBT11, MBM13], Mauderer et al. [MCNV14], Vinnikov et al. [VAF16] and Kramida et al. [Kra16].

In a first experiment on dynamic depth-of-field (DOF), Hillaire et al. test the effect for first person shooters [HLCC08]. The authors did not use real-time gaze tracking. Instead, the amount of DOF blur was derived from the circle of confusion (CoC) per-pixel based on the focus plane. The focal plane was estimated for an assumed salient, central area of the screen. Accommodation over time was simulated by temporal filtering of the focal distance. Half of the participants favored activating DOF rendering for gaming. Importantly, the rendered DOF effect did not decrease game performance.

Mantiuk et al. investigated gaze-point dependent DOF using eye tracking [MBT11, MBM13]. In the study, participants mostly reported a more natural feeling when using the gaze-contingent blur effect. However, the success of DOF rendering strongly depends on tracking accuracy [MBT11]. To improve tracking accuracy and stability for DOF rendering, the authors measured smooth pursuit eye movement and mapped the path onto estimated object motions in the scene. The object tracking approach resulted in a more successful DOF rendering in comparison to [MBT11].

The study of Mauderer et al. confirms previous findings with respect to perceived realism [MCNV14]. In addition, DOF rendering also results in better discrimination of object ordering. The effect can improve relative depth estimation. However, the accuracy of distance prediction is limited. The results of previous studies have been confirmed and extended in elaborate experiments by Vinnikov and colleagues [VAF16].

3.2.8 Dynamic Tone Mapping

Tone mapping is the process necessary to approximate the appearance of high-dynamic-range images on low-dynamic-range display devices or prints. A great deal of work has been done on tone mapping operators (TMO) for image and video processing. According to the categorization proposed by Eilertsen et al. [EUWM13], tone mapping operators aim to achieve different goals. Visual System Simulators (VSS) simulate limitations and properties of the HVS and try to derive a perceptually accurate reproduction of the captured or rendered scene. Scene Reproduction Operators (SRP) create the most perceptually faithful reproduction of color, contrast and sharpness, whereas Best Subjective Quality (BSQ) operators create the most preferred version with respect to subjective preferences or artistic goals. In the following, the most influential, perceptually-motivated VSS operators are briefly described. Detail information can be found in the survey papers by Eilertsen et al. and Fairchild on recent tone mapping operators [EUWM13, Fai15].

The TMO presented by Ferwerda et al. *globally* simulates eye adaptation over time and modulates visual acuity and color perception accordingly [FPSG96]. The model is tuned by psychophysical measurements. Pattanaik et al. uses exponential smoothing filtering for global temporal adaptation simulation, whereas different models are used for simulating cone and rod response [PTYG00]. Like Ferwerda et al. , Pattanaik et al. also use psychophysical measurements for calibration.

Ledda et al. propose a simple physiological model of eye adaption that approximates the *local* photoreceptor response with a temporally adjustable sigmoid curve. The time-dependent parameter is computed with respect to the characteristics of rods and cones, which results in a simulation of photopic, scotopic and mesopic vision conditions as well as receptor bleaching, and regeneration. . The luminance difference between succeeding frames determines the adaptation rate [LSC04].

Krawczyk et al. model temporal adaption using an exponential decay function [KMS05]. In addition, local contrast and optical aberrations are calculated by taking pupil size into account, resulting in naturally-looking scenes.

Benoit et al. present the Retina model TMO including a local, biologically-inspired retina model which enables spatiotemporal filtering with local cellular interactions and temporal stability [BAHLC09].

Mantiuk et al. propose the first real-time *gaze-dependent* tone mapping operator (GDTMO) by including eye tracking into the rendering pipeline [MM13]. The operator simulates eye adaptation based on fixation location. Temporal eye adaptation is controlled by the luminance of the gaze point area.

Recently, a complex GDTMO has been proposed by Jacobs et al. [EJGAC+15]. The operator simulates gaze-dependent global adaptation over time as well as a variety of secondary perception effects such as bleaching afterimages, mesopic hue shift, and desaturation under very dark and very bright illuminance conditions.

For evaluating the perceptual quality and fidelity of tone-mapped images and videos, image metrics inspired by perception such as the Structured Similarity Index (SSIM) or the High-dynamic Range Image Visual Difference Predictor (HDR-VDP) can be used [WBSS04, MDMS05, MKRH11]. The HDR-VDP is a metric that makes use of the described CSF variation to describe the perceived difference of two input HDR images. The HDR-VDP can be used for testing the perception of image compression distortions or the visibility of visual features. The predictor assumes local adaptation to luminance levels of a scene and filters the images using a normalized version of the CSF. The optical properties of the eye are also taken into account for the adaptation-dependent optical transfer function (OTF). After filtering both input images, the predictor splits the images into spatial and orientational channels. The sum of differences for all channels results in a final visual difference map. Tsai et al. introduce a foveated image quality metric that uses saliency maps [TL14]. The metric performs different measurements with respect to visual eccentricity. The foveal region adopts more strict quality assessment criteria than the peripheral image parts.

3.2.9 Gaze Guidance

Under normal circumstances attention is guided by visual features and the task of the user, which is exploited for passive gaze prediction. Strategies for gaze guidance are aiming for steering attention to a specified target location which can significantly differ from the natural fixation location. Therefore, gaze guidance requires altering the visible scene content.

Kosara et al. introduce the *semantic depth-of-field* for guidance based on the observation that gaze is attracted by high frequencies [KMH+02]. With this approach the target location is shown at high detail whereas non-salient parts are increasingly blurred towards the periphery.

Cole et al. and DeCarlo et al. both make use of stylized rendering algorithms for gaze guidance in images [DS02a, CDF+06]. In the first approach the target location is rendered with higher spatial detail than the non-salient image parts. The amount of detail is controlled by a perceptual model based on the contrast sensitivity function [DS02a]. Equivalently, in the second approach only the target gaze location is rendered at full quality. However, the remaining parts of the image are desaturated, blurred and faded out which reduces contrast in those parts [CDF+06].

Barth et al. enable gaze guidance for videos by augmenting the video with small bright red dots appearing at the target location [BDB+06]. The authors exploit the fact that sudden object onsets in the periphery attract attention. In 40% of the trials the appearing dots induce saccades towards the target. The stimulus was removed before the saccade was finished.

McNamara and Bailey introduced a more subtle, yet effective gaze guidance strategy [MBG08, BMSG09]. The authors apply image space modulations in the luminance channel to guide a viewer's

gaze through a scene without interrupting their visual experience. The principle has been successfully applied to increase search task performance as well as to direct gaze in narrative art [MBS+12]. Hence, the technique may support understanding of a painting in a gallery or a related use case but may also be useful for gaze guidance in simulators and training, pervasive advertising or perceptually adaptive rendering [BMSG09].

Pomarjanschiab and coauthors introduce gaze guidance in an automotive context [PDBB13]. An LED array was integrated into the interior of a car that enables an artificial moving light. In specified, potentially dangerous situations, such as driving off the street light of the LED array captures attention of the driver to guide the gaze back to the correct driving direction.

3.2.10 Perceptual Resolution Enhancement

Displaying high-resolution images on a low-resolution display is a sampling problem (Figure 3.3). For reconstruction, the high-resolution image is convolved with a spatial reconstruction filter for every output pixel. Well-known examples are the cubic splines derived by Mitchell and Netravali [MN88] and the Lanczos filter [Duc79]. In contrast, approaches for *perceptual resolution enhancement* rely on actively adapting spatial and temporal signal integration to go beyond physical pixel resolution.

Hara and Shiramatsu [HS00] inspected the influence of special pixel-color mosaics when moving an image at a specific velocity across the display but could not observe any improvement for the standard RGB layout. Similarly, *subpixel rendering* exploits knowledge about the arrangement of RGB color filters on a grid of photosensors for optimal filtering [Pla00] or masking defective subpixels [MK06]. The perceptual approaches by Didyk et al. [DER+10a] and Templin et al. [TDR+11] both take the HVS's smooth pursuit eye movement into account to display subimages at high refresh rates. The temporal integration in the human eye provides perceptually enhanced resolution. A similar approach is taken by Basu and Baudisch [BB09] who propose to move the image along a small circular path. However, circular motion has proven non-optimal [TDR+11]. While Didyk et al. [DER+10a] demonstrated the applicability of their approach only for linear motion, Templin et al. [TDR+11] transformed the image filtering into an optimization problem which is applicable to animated sequences.

Berthouzoz et al. [BF12a] combine apparent display resolution enhancement with super-resolution. A display-specific time-averaged point spread function is derived and used in the optimization process. Whereby the resolution enhancement method in [BF12a] yields comparable quality to [TDR+11], super-resolution enables enhancing the perceived spatial resolution of videos that already feature the display resolution [BF12a].

The achieveable quality enhancement of the mentioned techniques is affected by the apparent motion of the video content derived from the optical flow between the displayed frames. Recently, Wang et al. enable apparent resolution enhancement for near-eye light field displays for scenes that are ray-traced in real-time [WDZW15].

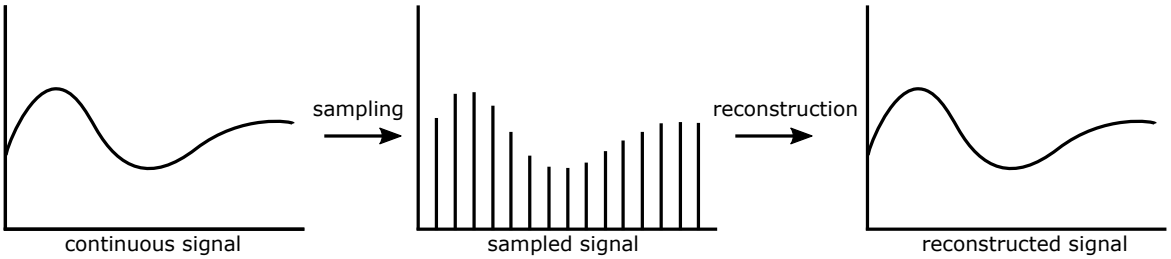

Fig. 3.3 (left) A continuous signal is sampled (center). The original signal can be reconstructed (right) if the sampling rate is sufficiently high according to the Nyquist-Shannon theorem [Sha49].

Berthouzoz and Fattal [BF12b] exploit the temporal summation effect of the eye to achieve resolution enhancement by small-amplitude vibrations of the display, synchronized with the screen refresh cycles. Damera-Venkata and Chang [DVC09] combine multiple display samples via the superimposition of image subframes from multiple projectors. This *display supersampling* can be optimized for antialiased and perceived super-resolution images. "Wobulated" projection use an opto-mechanical image shifter to slightly shift subimages for this purpose [AU05]. Unfortunately, these approaches require very specialized or calibrated hardware and are not directly applicable to off-the-shelf devices.

3.2.11 Gaze-contingent Video Filtering

Knowledge of where users will look can be beneficial, e.g., for video compression [DI03, HNA+11]. *Foveated video algorithms* have been proposed to deliver high-quality video at reduced bit rates by matching the compression rate to the acuity of the HVS. Geisler and colleagues propose low-bandwidth video communications based on real-time gaze data [GP98, GP99]. The approach uses a simple coding and decoding scheme based on an image pyramid which is generated from high-resolution video. Later, the authors extended the algorithm to compare video perception for people with normal vision and for patients with macular degeneration [PG02]. The authors used resolution maps to create videos in real time at variable resolutions across the visual field according to a given resolution function.

Lee et al. propose a rate control scheme for a foveated MPEG/H.263 video codec that is suitable for fast video rendering and for a given target bit rate [LPB98, LPB01]. An extension to passive gaze tracking is the multi-foveated MPEG compression scheme proposed by Dhavale et al. [DI03]. The approach uses saliency maps that combines bottom-up and top-down saliency features. Nikolov et al. make use of LODs derived from mip maps for generating bi-resolution and multi-resolution videos in real time [NNB+04].

Dorr et al. investigates the detection threshold of a gaze-contingent spatiotemporal filtering effect which removes frequencies from the video until the observer is able to detect visual differences [BDMB06]. The study confirms that the amount of spatiotemporal blurring that can be applied without being detected increases with eccentricity. The results also show that gaze-contingent tem-

poral filtering has an effect on the length of saccades which are shorter for the filtered videos. In a second study the authors test the algorithm with an eye-tracking head-mounted display [DBMB06]. Their video see-through HMD includes a pair of front cameras. The system performs gaze-contingent video manipulations on the visual input in real-time.

Ryoo et al. present a video streaming pipeline to save transmission bandwidth and processing ressources by downscaling a presented video in the peripheral area [RYS+16].

For quality assessment of foveated videos Wang et al. propose the foveated wavelet image quality index (FWQI) which models an eccentricity-based contrast sensitivity function [WBLK01]. Lee et al. introduce the foveal signal-to-noise ratio (FSNR) as an objective quality criterion to measure foveated image/video quality against compression gain [LPB02]. Rimac et al. introduce the foveated mean squared error (FMSE) which includes peripheral acuity reduction and motion dependency of the acuity function [RDVŽ10].

Further Reading

More information on gaze-contingent displays, perceptual rendering strategies and acceleration techniques is provided in Refs. [Lue03, RWPD10, DGY07, Tin14, HVDFF14, BS14].

Chapter 4

Apparent Display Resolution Enhancement for Arbitrary Videos

Contents

Display resolution is frequently exceeded by available image resolution. Recently, apparent display resolution enhancement techniques (ADRE) have demonstrated how characteristics of the human visual system can be exploited to provide super-resolution on high refresh rate displays [DER+10a, TDR+11]. This chapter adresses the problem of generalizing the apparent display resolution enhancement technique to conventional videos of arbitrary content. A novel optimization-based approach is derived to continuously translate video frames in such a way that the added motion enables apparent resolution enhancement for the salient image region. The optimization takes the optimal velocity, smoothness and similarity into account to compute an appropriate trajectory. In addition, an intuitive user interface is provided which allows one to guide the algorithm interactively and to preserve certain artistic camera motions. The approach is evaluated in a perceptual study. The results verify high apparent rendering quality and demonstrate the versatility of the proposed method for a variety of test scenes.

4.1 Introduction

Modern cameras and rendering hardware are able to produce highly detailed images. Sophisticated tone and gamut mapping algorithms adapt them to available display capabilities. Even though hardware constantly evolves, limitations in color, luminance, and spatial resolution constrain the range of reproducible images on various devices. Latest advancements such as apparent image contrast [PSL99] or apparent brightness [YIMS08] have shown that it is possible to go beyond the physical limitations of display devices by exploiting characteristics of the human visual system (HVS).

This work addresses the problem of apparent spatial resolution enhancement. High-definition TVs and projectors have become ubiquitous, but the resolution of current digital cameras and cinema movies is still up to one order of magnitude higher than what these displays can currently show. The necessary downsampling procedure results in the loss of fine details, such as fur, hair or general high-frequency image features. On the other hand, the refresh rate of commodity TVs and projectors increases more and more and 120 Hz TVs are available today. With active-matrix organic light-emitting diode (AMOLED) technology even higher refresh rates ($> 1000\,\text{Hz}$) can be achieved and will be available in the near future. The challenge is how to provide a better viewing experience given available high-resolution image data and limited display resolution.

It has been shown by Didyk et al. that the integration on the retina of moving high frame-rate, low-resolution subimages results in an increased perceived resolution, if displayed above the *critical flicker frequency* (*see* Chapter 2.3) [DER+10a].The necessary smooth pursuit eye movement (SPEM, *see* Chapter 2.4) is induced by artificially moving the static image at constant velocity. Templin et al. have shown that a similar effect can be achieved by exploiting the natural movement in high frame-rate videos [TDR+11]. Berthouzoz et al. additionally integrate a super-resolution technique into the optimization process to achieve resolution enhancement in comparable quality also for input videos that are recorded with display resolution [BF12a].

The apparent display resolution enhancement (ADRE) technique is affected by several aspects: first, to perceive high contrast, high refresh rates are necessary [KL07]. Second, for best perceived spatial resolution, the movement of the displayed video needs to be along one of the four diagonals and at a specific velocity [DER+10a]. The more the movement differs from these requirements the less pronounced the effect will be.

The approach described in this chapter extends the work of Didyk et al. [DER+10a] and Templin et al. [TDR+11] in several important aspects (Fig. 4.1). It is shown how slight changes to a standard high-resolution, low frame-rate video can support ADRE to perceive a higher resolution. A stronger diagonal movement is enforced at the required speed by computing the flow of the most salient regions in the video and by shifting the video content along an optimized trajectory. Specific attention is paid to subtle and smooth changes to incorporate original movements in the video. The optimization is based on an energy minimization which incorporates saliency (to focus the optimization on regions of interest), smoothness (to support SPEM and prevent flickering), similarity to the original footage (to prevent the movement from going astray) and resemblance to the optimal direction and velocity (to provide the best possible input to the ADRE algorithm). In addition, to handle low frame-rate videos, a motion path is computed to offset duplicated frames to further support apparent resolution enhancement. A specialized user interface allows one to interactively change the optimization parameters within the video for artistic guidance of the optimization. In contrast to [TDR+11] and [BF12a], the proposed approach enables apparent resolution enhancement even for scenes that do not contain any movement and for which typical optical flow computations are difficult or impossible.

One possible application of the presented approach are common high refresh rate TVs and projectors to display high-resolution videos. Furthermore, the approach could be used to reproduce high-resolution Virtual Reality video playback on currently rather low-resolution VR displays. When watching a video on a display with a different aspect ratio, e.g. 4:3 instead of 16:9, the cropping area can be optimized to support ADRE.

The remainder of this chapter is structured as follows: In Section 4.2 background information on the general apparent resolution model is provided. Inherent problems for traditional videos are described in Section 4.3. The extended model is presented in Section 4.4 and the two-stage saliency scheme in Section 4.5. The application of this extended model provides important input to a novel trajectory optimization algorithm in Section 4.6. In Section 4.7 a specialized user interface is described that allows to preserve certain artistic camera motions or to manually restrict the optimization. Conducted perceptual experiments and user studies are analyzed in Section 4.8 and discussed in Section 4.9, before the chapter concludes in Section 4.10.

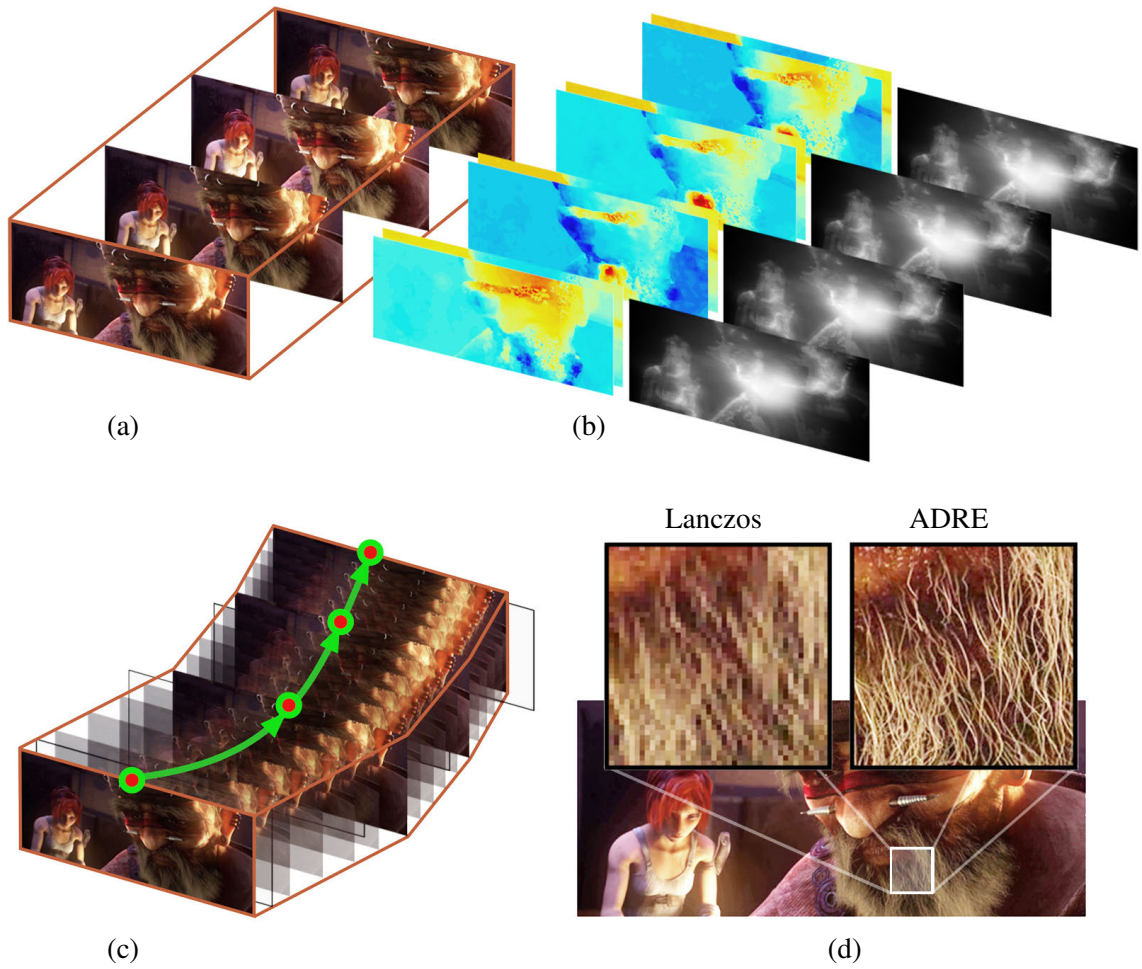

Fig. 4.1 Improving apparent display resolution enhancement for general video footage.
Given a standard 24–30 Hz video (a) optical flow and importance maps are computed (b) to temporally upsample and offset the video along an optimized smooth trajectory (c). This results in increased perceptual resolution (d) exceeding the physical resolution of the display using apparent display resolution enhancement algorithms (ADRE).

4.2 Apparent Display Resolution Enhancement

The model of Didyk et al. describes the response r of a receptor in the human eye as the integrated signal of the observed intensity $I(t)$ over a time T [DER+10a]. When an observer focuses on a detail in a moving image or video, the eye tries to follow its trajectory (smooth pursuit eye motion, SPEM). If the receptor moves along a smooth path $p(t)$ the integrated result is:

$$r(I, p(.)) = \int_0^T I(p(t), t)dt \tag{4.1}$$

Thus, intensities of neigboring pixels are perceptually mixed when the path crosses pixel boundaries. Owing to this movement, as well as the higher density of photoreceptors on the retina in comparison to screen resolution, neighboring receptors may reveal a different solution to the integral (*hold-type blur*). Equation (4.1) does not hold in general [vH05], although it is a valid assumption for signals displayed above the *critical flicker frequency* when pixel intensities over time are fused into a steady appearance [KL07]. Since I is a discrete function in space (pixels) and time (frames), Equation (4.1) can be reformulated as

$$r(I, p(.)) = \int_0^T I(p(t), t)dt = \sum_{i,j,k} w_{i,j,k} I^k_{i,j} \ , \tag{4.2}$$

where

$$w_{i,j,k} = \int \chi_{i,j}(p(t))\chi_k(t)dt \ . \tag{4.3}$$

The characteristic function $\chi_{i,j}(p(t))$ equals one if $p(t)$ lies within the pixel (i, j) and zero otherwise; $\chi_k(t)$ is a similar function for time interval k, i.e. frames. The weight $w_{i,j,k}$ is normalized by the total length of the path $|p|$. Utilizing the hold-type blur induced by SPEM in combination with high refresh rate screens results in apparent display resolution enhancement [DER+10a]. The intent is to optimize the subimages so that

$$\mathbf{W} \begin{pmatrix} I^1_L \\ \vdots \\ I^k_L \end{pmatrix} - I_H = 0 \ , \tag{4.4}$$

where I^i_L is the i-th low-resolution subimage and I_H is the original high-resolution image. The underlying assumption is that there is a one-to-one mapping between receptors and pixels in the high-resolution image so that $r_{x,y}$ is close to $I_H(x, y)$, i.e. one row in I^k_L describes the path of one receptor along the subimages.

If the subframes are displayed at high refresh rates, intensity integration on the retina reconstructs high frequency details. A-priori assumption about SPEM yields $\mathbf{W}$. The sparse matrix $\mathbf{W}$ is computed, using only the respective receptor's starting position and motion.

Templin et al. [TDR+11] extend this model to videos by approximating the complex motion in an animation with many simple integer motions computed for every possible triplet of frames, i.e. for subframes {1,2,3}, {2,3,4} etc.

$$\mathbf{W}\left(\mathbf{I_L} \right) - \mathbf{I_H} = 0, \tag{4.5}$$

where $\mathbf{I_L}$ is the vector of all subframes and $\mathbf{I_H}$ the vector of the original high-resolution images. The triplets are encoded in the appropriate weighting matrix $\mathbf{W}$. its optical flow. The equation system is usually overdetermined so that no solution exist. Hence, subimages $\mathbf{I_L}$ are derived by solving the system as a constrained quadratic minimization problem [TDR+11].

4.3 Problem Statement

Unfortunately, for general videos no sufficiently accurate solution to Equation (4.5) may exist. Between eye saccades, the foveal area of the eye may follow any feature f in the video, and the path $p(t)$ is dependent on the movement of f. Several cases exist where ADRE fails, Fig. 4.2. If f does not move at all or is too slow (r_1, orange), the integrated signal of neighboring receptors seeing the same low-resolution pixel all perceive the same stimulus, Equation (4.2). Thus, no resolution enhancement is possible. On the other hand, if f moves faster than $2.5\,{}^{\mathrm{deg}}/_{\mathrm{s}}$ (r_2, blue), stabilization of the image on the retina cannot be guaranteed anymore [LRP+06]. In the case of horizontal or vertical movement (r_3, red), resolution enhancement is only possible in horizontal or vertical direction, respectively. Owing to reaction times of the HVS, the eye cannot follow sudden kinks in the movement of f (r_4, green). Optimal ADRE is achieved only if f moves smoothly along the diagonal at the speed of one high-resolution pixel per subframe in x and y direction (r_5, magenta).

4.4 Extended ADRE Model

The residual error e from minimizing Equation (4.5) is an objective measure of the quality of the ADRE algorithm given the assumption of an aliasing-free original image and perfectly reconstructed receptor movements, respectively *optical flow*.

Optical flow is described as the apparent velocities of brightness patterns in an image [HS81]. Templin et al. assume the receptor motion to match optical flow in the animation [TDR+11]. This assumption, however, does not hold for general videos. In addition, as optical flow follows feature motion, flow direction and speed usually may be non-optimal for ADRE (Sec. 4.3).

Differing from the algorithm of Templin et al. the input video frames $\mathbf{I_H}$ are translated by $\mathbf{T}$ in order to improve the result of ADRE:

$$\mathbf{W}\left(\mathbf{I_L} \right) - \mathbf{T}(\mathbf{I_H}) = 0 \ . \tag{4.6}$$

Note that in this case $\mathbf{W}$ and $\mathbf{T}$ are dependent variables because a change of $\mathbf{T}$ also changes the optical flow and therefore $\mathbf{W}$. Restricting $\mathbf{T}$ to a discrete translation for each frame prevents resampling of

the image which would otherwise annihilate most of the ADRE effect and render it useless. The operator T^k describes the absolute translation of the k^{th} frame I_H^k of the input video. For simplicity of explanation, T^k is used as both the translation function and the corresponding displacement vector.

4.5 Saliency Model

ADRE algorithms are based on the assumption that the receptors of the eye follow the optical flow in the video [TDR+11]. The movement of the human eye however, has only two degrees of freedom. If the optical flow is non-constant across the image, e.g. different foreground and background movement, the integration result of the receptors is not in accordance with the optimization in Equation (4.5). It further implicates that in some cases no sufficient translation $\mathbf{T}$ can be found as the required changes may cancel each other out. Assuming that the eye movement is known, it is a valid simplification to optimize $\mathbf{T}$ only for receptors of the fovea due to the rapid falloff in acuity away from the foveal region [CSKH90]. The proposed approach uses image saliency to model eye fixations. In our two-component saliency model, a saliency map S^i is computed from both objectives, automatic saliency measures and eye tracking data from a user study for each frame I_H^i of $\mathbf{I_H}$. Using stand-alone saliency metrics turned out to be insufficient compared to results including eye tracking data. As suggested by previous literature [JDT12], the proposed approach combines averaged ground truth data from eye tracking with an objective saliency approach to allow for robust and accurate gaze prediction. If the quality of future saliency algorithms increases a purely software-based solution for gaze prediction may become possible. However, optimizing automatic saliency generation is beyond the scope of this thesis.

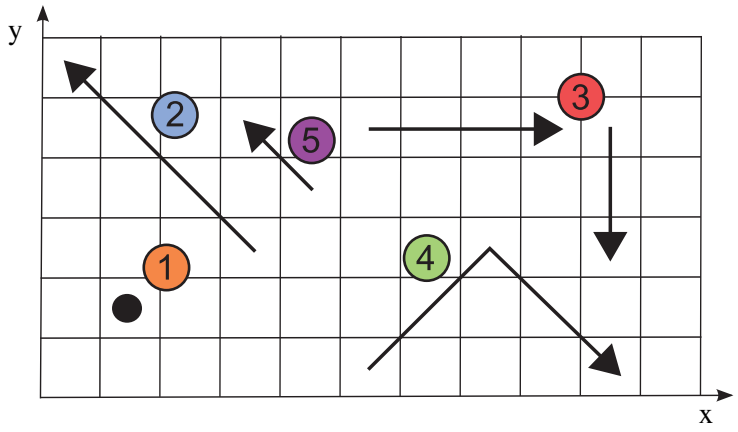

Fig. 4.2 Receptor motion. By projecting images at high frame rate, a receptor is apparently moving in 2D space (velocity is constant along the depicted arrows). Different failure cases for ADRE can occur: Receptor r_1 (1, orange) has no or a too small velocity to achieve resolution enhancement; r_2 (2, blue) is too fast; for r_3 (3, red) resolution enhancement occurs only along the horizontal or vertical axis; the movement of r_4 (4, green) is optimal in direction but cannot be followed by SPEM; r_5 (5, magenta) shows a desirable movement for ADRE.

4.5.1 Subjective Saliency

In all conducted eye tracking experiments an EyeLink 1000 eye tracker from SR Research has been used [Res16]. While subjects watched the videos, their gaze paths (relative position within the video) were recorded at 240 Hz. 17 subjects with an average age of 25 and normal or corrected-to-normal vision participated in the perceptual study for saliency generation. To extract the salient regions in the videos, fixation points of the subjects were recorded, consisting of an *x*-*y* coordinate on the screen and the duration in milliseconds.

The output from the conducted eye tracking experiments is a single position vector for each frame per video and participant. The assumption is that in the limit, i.e. with an infinite number of participants watching the animation for the first time, the normalized sum of the eye tracking data is the true saliency function $\mathbb{S}$. Hence, the data from the eye tracking experiments is a sparse sampling of $\mathbb{S}$, and estimating $\mathbb{S}$ becomes a reconstruction problem. Due to the restrictions on $\mathbf{T}$, the images were downsampled by three octaves, smoothed with a Gaussian filter of standard deviation $\sigma = 10$, and each resulting saliency map S_E of each frame was normalized. In the experiments around ten participants were sufficient because variance in gaze among all subjects was low. This result confirms findings of a related benchmark provided by Judd et al. [JDT12].

4.5.2 Objective Saliency Features

For automated saliency estimation the approach by Cerf et al. [CHH+09] is applied that uses a combination of low-level features and high-level semantics. The low-level features are based on contrast of color, intensity and orientations in the image [IKN98]. For the high-level semantics, face detection [VJ04] and person detection [FMR08] was employed as humans are usually attracted to faces and people [Gol13]. The detector of Viola and Jones proved to work fast and sufficiently robust for the tested scenes [VJ04]. Of course, any other saliency detector could be used instead, e.g. [ZR12]. The result is saved in an intermediate saliency map S_O^i.

The final saliency map S^i is derived for each frame as a weighted average of measured saliency S_E^i, predicted saliency S_O^i, and a constant λ_s by the equation

$$S^i = \left((1-\lambda_s) + \lambda_s \cdot \left(\alpha S_O^i + (1-\alpha) S_E^i\right)\right). \tag{4.7}$$

The constant $\lambda_\mathbf{s} \in [0,1]$ steers the influence of non-salient regions. This is important in scenes where the foreground is moving fast and suffers from motion blur but the background contains fine details. In our experiments $\lambda_\mathbf{s}$ was set to 0.25. α should be chosen depending on the reliability of the measured saliency S_E. In the conducted experiments α was also set to 0.25.

The saliency map is thresholded before being used in the optimization step described in the next section. This approach delivered sufficient maps in all of the test cases. However, this approach can be costly and invasive. Instead of using measured gaze data, semi-automatically created saliency maps,

e.g. using Adobe After Effect's Rotobrush™, can be employed. A coarse estimate of S is generally sufficient for the proposed algorithm.

(a) Original motion (Optical Flow) **(b)** Optimized relative motion

Fig. 4.3 Salient region motion optimization. In the salient region (motorbike driver) the optical flow shows only horizontal movement (a). In this case ADRE can be applied in a single direction only. After optimization (b) the salient region moves diagonally across the screen, ideal for ADRE.

4.6 Trajectory Optimization

It is assumed that a sufficient condition for $\mathbf{T}$ to serve as a good transformation for ADRE is given by four essential conditions:

1. Similarity to the optimal input for ADRE, which is a one pixel flow along one of the diagonals per frame of the high-resolution video [DER+10a];
2. Smoothness of directional change;
3. Similarity to motion in the original footage;
4. Visibility of image regions steered by saliency.

An explanatory example of the optimization is given in Fig. 4.3. The fourth condition simplifies optimization given the assumption that the flow inside the salient regions does not diverge. Let $\mathbf{u}^k$ be the optical flow from the original video $\mathbf{I_H}$ from frame k to $k+1$. Instead of evaluating and optimizing for every pixel of the video, the mean flow μ weighted by the saliency is used:

$$\mu^k = \sum_{i,j} S^k(i,j)\, \mathbf{u}^k(i,j), \tag{4.8}$$

The cumuluative sum of μ describes the trajectory of the salient region in $\mathbf{I_H}$.

Let $\mathbf{v}^k$ be the synthetically added translation we search for from frame k to frame $k+1$. The following energy terms incorporate the frame-dependent weights $w^k_{\text{vel}} \in [0,1]$, $w^k_{\text{smooth}} \in [0,1]$ and $w^k_{\text{imp}} \in [0,1]$, which can be adjusted by our user interface described in Section 4.7.

The first condition is formulated as the difference of the newly synthesized flow from the optimal diagonal movement $\mathbf{v}_{\text{opt}}$:

$$E_{\text{vel}} = \sum_{k=1}^{n} w^k_{\text{vel}} \, \|\mu^k + \mathbf{v}^k - \mathbf{v}^k_{\text{opt}}\|_2^2 \tag{4.9}$$

Further, smoothness of the synthetic trajectory is enforced by minimizing the norm of the numerical derivative of the artificial flow:

$$E_{\text{smooth}} = \sum_{k=1}^{n-1} w^k_{\text{smooth}} \, \|\mathbf{v}^{k+1} - \mathbf{v}^k\|_2^2 \tag{4.10}$$

Finally, the translated video is prevented from drifting too far out of the original viewport by defining an additional energy term E_{imp}. For this, a distance map D^k is derived from the saliency maps S^k first. Each pixel in D^k stores its distance to the closest salient pixel. Then, D^k is transformed into an *importance map* V^k by mapping it to the range $[0,1]$ using $V^k(i,j) = 1 - D^k(i,j)/\max(D^k) + \varepsilon$, with $\varepsilon > 0$. A visualized example of the importance maps is given in Fig. 4.1. E_{imp} assures that the salient region S^k of each video frame stays within the original viewport V after applying the translation T^k. Additionally, the term generally penalizes large translations.

$$E_{\text{imp}} = \sum_{k=1}^{n} w^k_{\text{imp}} \, (\sum_{(i,j)} V^k(i,j) - \sum_{(i,j) \in (VP \cap T^k(VP))} V^k(i,j)), \tag{4.11}$$

E_{imp} computes the sum of all values in V in the transformed video outside the original viewport VP.

The final energy term is a weighted sum of its subterms:

$$E = \alpha \, E_{\text{vel}} + \beta \, E_{\text{smooth}} + \gamma \, E_{\text{imp}}. \tag{4.12}$$

The weighting factors α, β and γ are chosen suitably for the coarse scale of the individual energy terms. For the test scenes, $\alpha = 10^3$, $\beta = 10^{-3}$ and $\gamma = 10^4$ are used.

The above formulation is used to perform an Expectation Maximization-like optimization by iteratively refining an initial zero vector $\mathbf{v} \in \mathbb{R}^{n \times 2}$. Each iteration implicates alternating between finding an optimal movement $\mathbf{v}_{\text{opt}}$ for each frame in Equation (4.9), and updating $\mathbf{v}$ to minimize Equation (4.12). Note that at this stage $\mathbf{v}$ is treated as a vector field. Since $\mathbf{v}$ changes in this step, the classification of $\mathbf{v}_{\text{opt}}$ for each frame may also change. Therefore, the two steps are repeated until convergence, which is guaranteed as each step minimizes Equation (4.12) further. The mean flow μ in Equation (4.9) assures that the natural movement in the video is taken into account during optimization. In the expectation step, the optimal flow is selected from

$$\mathbf{v}^k_{\text{opt}} \in \left\{ (1,1)^\top, (1,-1)^\top, (-1,1)^\top, (-1,-1)^\top \right\}$$

for each frame k in order to minimize Equation (4.9). To prevent jittering of $\mathbf{v}_{\text{opt}}$, the optimal flow $\mathbf{v}_{\text{opt}}$ is kept constant for c frames. Therefore, bundles of c frames are successively created, and the mean direction of the salient region is computed, i.e. $\frac{1}{c}\sum_{i=1}^{c}(\mu^{k+i}+\mathbf{v}^{k+i})$. Out of the four possibilities of $\mathbf{v}_{\text{opt}}$ the value closest to the mean direction is selected. c is a user defined variable. To further enforce smoothness of the overall trajectory of the salient region, an additional diffusion step to the calculated $\mathbf{v}_{\text{opt}}$ is applied by averaging each $\mathbf{v}_{\text{opt}}^{k}$ with its neighbors and repeating the process for m iterations. Per default we set $m = 5$.

In the maximization step, a non-smooth numerical optimization method is employed to minimize Equation (4.12) w.r.t. $\mathbf{v}$ [Ove10]. Recall from definition (4.11) that E_{imp} is not smooth with respect to $\mathbf{v}$. For fast convergence of the optimization algorithm, the exact gradient is necessary. The algorithm usually converges after two to four EM-steps. An example is given in Fig. 4.4.

Finally we obtain $\mathbf{v}^k + \mathbf{u}^k$ as the ADRE-optimized optical flow.

4.6.1 Temporal Upsampling

Due to the critical flicker frequency, best temporal contrast is perceived for 120 Hz animations [DER+10a, KL07]. Unfortunately, standard movies are captured at a much lower frame rate, usually 24 or 30 Hz. Simple duplication of the video frames is no recommendable solution as the flow between intermediate frames will be zero, in which case the ADRE algorithm cannot produce an enhanced output. Image interpolation algorithms would require prohibitive resampling. To compensate for this, first, the optimized trajectory is computed based on the original video, but the magnitude of the optimal flow $\mathbf{v}_{\text{opt}}$ and the initialization of $\mathbf{v}$ is multiplied by a factor M, which is four or five for 30 Hz and 24 Hz recording frame rate, respectively. Each image is then duplicated $M-1$ times, and the m-th entity $I^{k,m}$, $m \in \{0,\dots M-1\}$, of image I^k is translated according to:

$$T^{k,m} = T^k + \text{round}\left(\frac{m}{M}(\mathbf{v}^k + \mu^k)\right). \tag{4.13}$$

This translation supports ADRE as it smoothes the movement of the salient region for the duplicated frames. In addition, this step gives the exact ground truth flow for each duplicated image from Equation (4.13).

Computing the correct optical flow for the original images can be delicate in complex scenes. A wrong flow can result in severe artifacts when using ADRE. If the results show that no sufficiently correct optical flow can be computed, only the displacement from Equation (4.13) is used as input to the ADRE algorithm by Templin et al. [TDR+11].

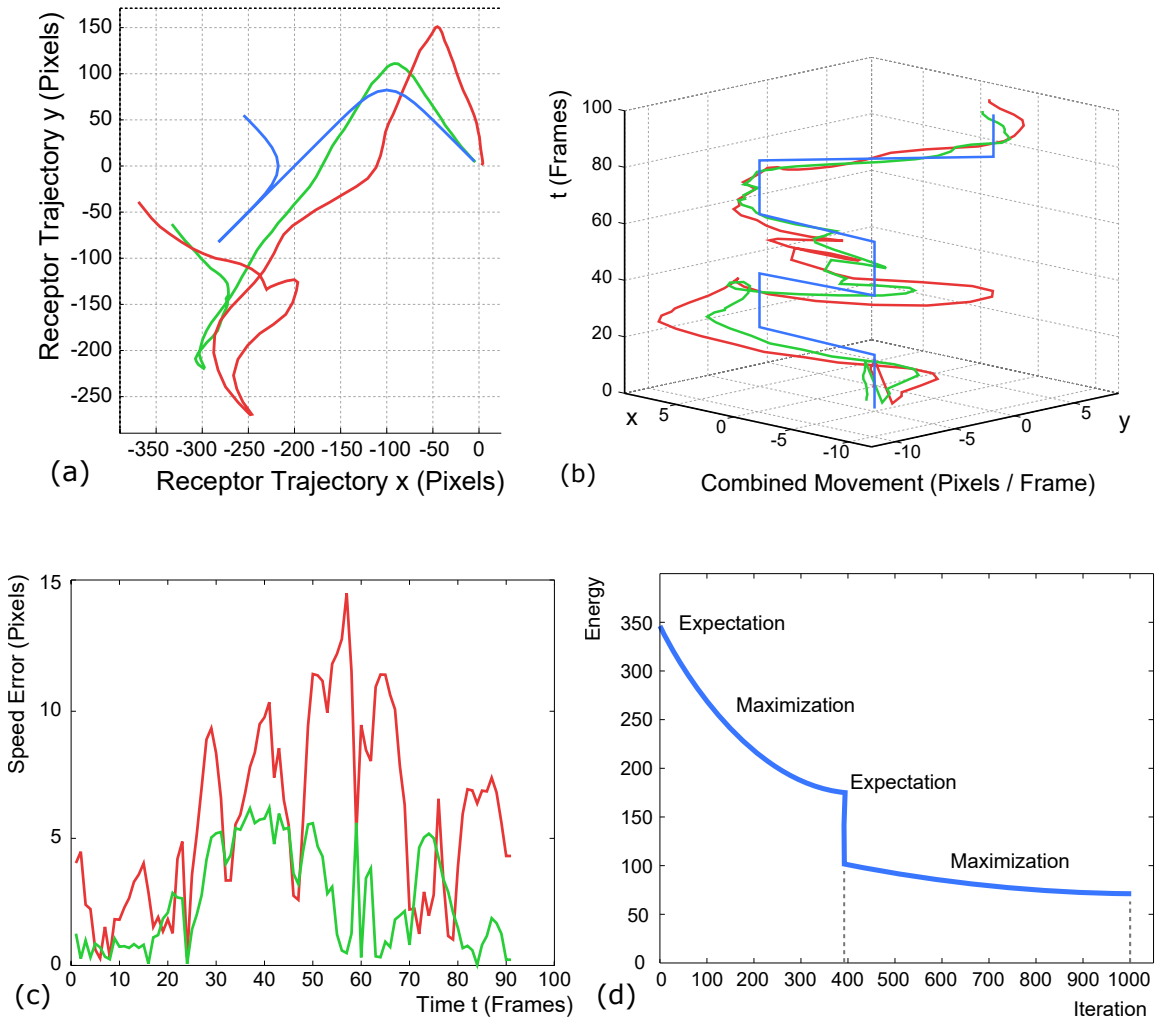

Fig. 4.4 Trajectory optimization for the scene EXPRESS. (a,b) Visualization of the cumulated velocities and velocities as a function of space (x,y displacement) and time (frames). The optimized trajectory (green) still resembles the original trajectory (red) but is closer to $\mathbf{v}_{\text{opt}}$ (blue) for optimized ADRE support. The magnitude of the optimal velocity $\mathbf{v}_{\text{opt}}$ in the high-resolution 24 Hz video is $\sqrt{5^2+5^2}$ pixels. The vertical and horizontal transitions of 5 pixels result in the optimal pixel motion velocity of 1 pixel per frame in the optimized 120 Hz video. (c) The red curve describes the deviation of the mean velocity μ from $\mathbf{v}_{\text{opt}}$ in the original video. The green curve describes the deviation after applying our optimization. Strong peaks give evidence of a change in trajectory. (d) Plot of the energy level as a function of maximization iterations.

4.7 User Interface Layout

An editor provided with the gaze-contingent video resampling approach lets the user interactively modify the video trajectory by steering the optimization. This becomes necessary if the computed trajectory violates artistic intentions. A screenshot is shown in Fig. 4.5.

The components of the interface consist of a video preview window and an editing area. The video view shows the transformed video itself, the saliency map as an optional overlay (shown in yellow) and the action-safe frame (yellow) and title-safe frame (red), which both are standardized values in movie production [EE16]. The action-safe and title-safe areas define margins to the four edges of the frame, so that all essential action and text are protected on any display type. Using this view, the user can easily analyze the synthetic motion and its magnitude. The video is cut into separate shots which are optimized independently to avoid jittering artifacts at shot boundaries.

The editing window is subdivided into a navigation pane for playback and several panels below which can be blended in or out as desired. Each one is showing one of the three energy terms influencing the optimization. At the top of each panel the user can specify keyframe values to set the relative influence of each parameter and energy weighting term throughout the video. To the left of each of the error panels the weighting factor with regard to Equation (4.12) is set. All time-dependent parameters are linearly interpolated between each pair of keyframes. At the bottom the relative error of each energy function is plotted with a rainbow color scale. This gives the user direct visual feedback on how any editing changes influence the quality of the later ADRE. Finally, the velocity control pane additionally contains three plots of the original velocity of the importance region (blue line), the optimized velocity (green line) and the theoretically optimal velocity (red dashed line). For example, in case a certain camera movement is essential in parts of the video, the user simply increases the influence of the importance map term (‘Visibility Control’) for these frames, and the optimizer adjusts the trajectory accordingly to closer follow the original motion path.

Fig. 4.5 Screenshot of the interactive ADRE editor. A preview of the modified video is shown in the top-right area. The video can be blended with the saliency maps (yellow). The bottom area shows a timeline of the video and a visualization of the different error terms. The user can interactively control the induced motion by setting keyframes and steering the influence of each term throughout the video. Error plots and trajectory for a new configuration are updated interactively.

Fig. 4.6 Example frames of scenes used in the perceptual study. Shown here are LUPE, SINTEL, GARDEN and BIG BUCK BUNNY. *Image courtesy of* Stephen Higgins, Evin Grant and the Blender Foundation.

4.8 Experiments and Results

For a 30-second, 24 Hz, 4K video the trajectory optimization implemented in MATLAB and C++ converges in one to five seconds, enabling interactive adjustment of the trajectory. The computational complexity of the proposed technique is dominated by the ADRE algorithm by Templin et al. . Because of the sheer size of the video 4k footage which does not fit into GPU memory a multi-threaded CPU version has been used for video resampling. Computation takes around 30–45 seconds per subframe at 4K resolution on an Intel i7-960 with 3.2 GHz and 12 GB of RAM. For quality analysis of the technique, two test data sets have been created where different goals have been pursued. The first data set (Maximum ADRE) contains 12 videos and has been optimized for best ADRE effect and lower-ranked similarity to the original video. The second video data set (Minimum change) with eight video sequences has been optimized for close similarity of the synthetic trajectory to the original movement.

4.8.1 Objective Enhancement – Statistics

The residuum of Equation 4.5 is an objective quality measure of the downsampling procedure, assuming perfect SPEM. The results for the original video and the optimized video as input have been compared to the algorithm of Templin et al. [TDR+11]. The proposed technique achieves an

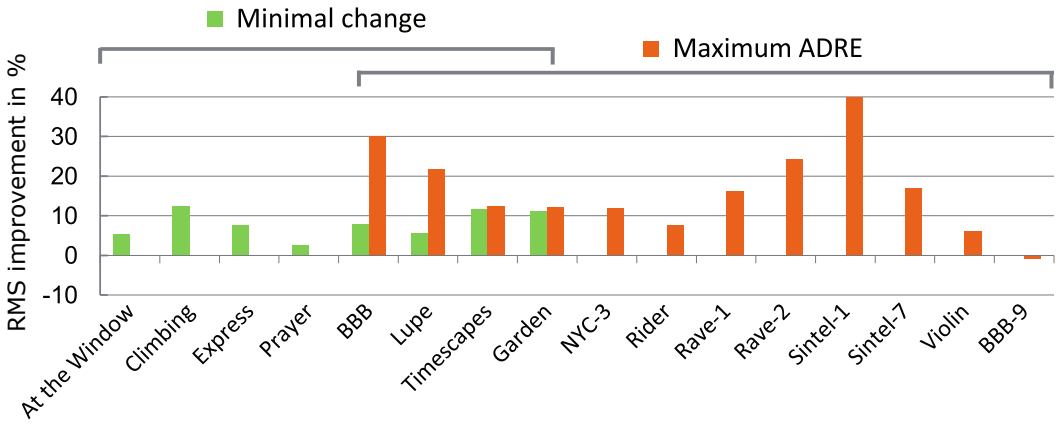

Fig. 4.7 Quantitative image improvement. The plot shows the improvement of the root-mean-squared-error using the optimized video variant as input to Templin's algorithm [TDR+11] in comparison to using only the original video (default). Higher percentage values indicate better results. One dataset (orange) is optimized for best ADRE effect whereas for the second (green), a trade-off between ADRE effect and minimal additional movement is targeted.

improvement of the root mean square error (RMSE), which computes the difference of the high resolution frames and the perceived downsampled images of 17%, on average, for the Maximum ADRE data set, and 8 % for the Minimum change data set, Fig. 4.7. Videos of the second data set show less improvement for ADRE since the manipulated motion is kept close to the original one. Note that the overall RMSE can also increase in non-salient regions, e.g. in the BIG BUCK BUNNY (BBB) scene as the proposed technique optimizes locally for the salient regions. However, this happened only in a single scene, and the increase in RMSE was below 1%.

4.8.2 Subjective Enhancement – Perceptual Study

To further validate the effectiveness of the approach a perceptual study with 21 participants has been conducted for both video data sets. All participants had normal or corrected-to-normal vision. They had no previous knowledge about the goal of the project nor the technique used. The subjects were seated in front of the monitor in a semidark room. They had been instructed orally regarding the procedure of the experiment. There was no time limit for solving the tasks.

The aim of the study was to show that the proposed method outperforms previous downsampling approaches for typical 24–30 Hz videos and that the editing tool enables controlling the trade-off between noticeability of the applied transformation and effectiveness of the approach. The novel method has been compared to static Lanczos downsampling and to the ADRE technique of Templin et al. [TDR+11] which directly uses optical flow as the predicted retinal path. As discussed in [BF12a] the quality of resolution enhancement achieved by the technique of Berthouzoz et al. is comparable to

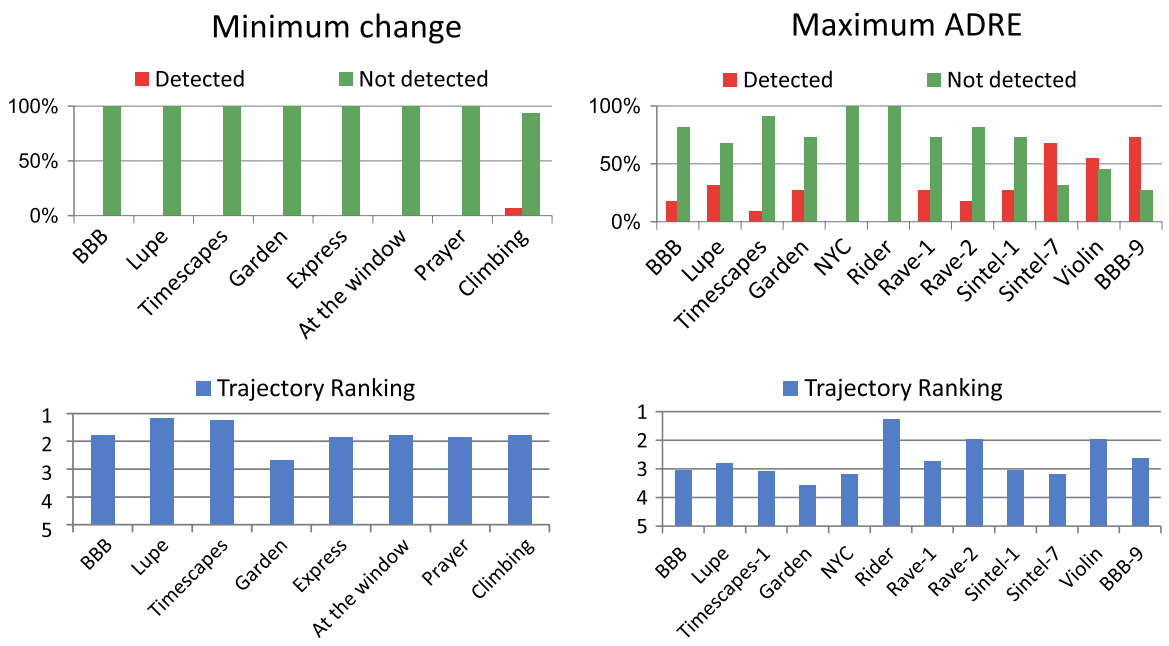

Fig. 4.8 Trajectory conspicuity. (first row) After viewing each video for the first time, participants stated if they had detected anything conspicuous regarding the video. (second row) After viewing the video several times they were asked to rank the annoyance of the trajectory between 1 (pleasing), 3 (neutral) and 5 (disturbing).

the results of Templin et al. for the case of a high resolution video downscaled to the display resolution. Hence, no further comparisons to [BF12a] have been made.

As display, a 23.6 inch (diagonal) 120 Hz Acer GD245HQ monitor was used at a resolution of 1920×1080 pixels. The subjects viewed the monitor orthogonally at a distance of 20 inches. A 120 Hz frame refresh rate was used. For the Lanczos reconstruction filter and Templin's algorithm the original videos have been temporally upscaled to 120 Hz by frame duplication. Hence, the optical flow is zero for all pixels between the duplicates. Note that using the original video for Templin's algorithm is not recommendable as the integration time is too long with standard videos.

In the perceptual study several test scenes were considered, including varying types of motion (weak/strong), styles (real-world footage and animations), ambiguities in motion and saliency, and transparency (which is a still unsolved challenge for optical flow in natural scenes). Example frames are shown in Fig. 4.6.

Motion perception In the first part of the study the conspicuousness of the video frame translation was analyzed. The modified 120 Hz videos were presented to the first-time viewers without instructing them about the technique. The participants were then asked whether they noticed anything unnatural in the videos and if so what it was.

The analysis shows that less than a third of the users noticed modification in the videos optimized for ADRE (Fig. 4.8, max. ADRE). In these cases the original video material hardly contained any motion as a stylistic intent (SINTEL-7, VIOLIN, BBB-9). In all other cases the optimization was able to produce a trajectory that was unsuspicious for first-time viewers. For the second data set manipulation was subtle so that almost no subject noticed a modification (Fig. 4.8, min. change).

Richness of Detail In the second part of the study the proposed method was compared with Lanczos filtered videos (with lobe parameters 4,5 and 6) and Templin's ADRE [TDR+11] applied to the original 120 Hz video. The required optical flow for Templin was computed using the algorithm of Werlberger et al. [WTP+09] with an illumination corrected intensity term.

In pairwise comparisons, the participants had to decide which video preserved more details. Results are given in Fig. 4.9. The videos were shown in a loop. The novel approach was shown either on the left or right side. When optimized for minimum change of the original camera path the new approach is rated slightly better in terms of detail reconstruction for most scenes. However, significance to differences are observed only for three videos (TIMESCAPES, EXPRESS, PRAYER) for which optimal movement can be achieved. For the videos of the second data set, participants judged the new technique as significantly better. The results show that the degree of permitted manipulation strongly affects the perceivable improvement in apparent resolution enhancement. Compared to Templin the novel approach performed better in videos which originally contained little motion or non-diagonal motion direction (GARDEN, LUPE, PRAYER, BBB). This significance is statistically validated using a χ^2 test, successfully falsifying the null hypothesis. Proving that in most cases there is a noticeable quality improvement with the novel approach. The proposed algorithm was judged even higher in scenes where the RMSE globally increased (BBB-9).

Proximity to original movement The results show that the movements are noticable in case of adding motion to still shots when optimizing for maximum ADRE. However, the movements have not been rated as being disturbing. This shows that the developed editing tool enables one to successfully avoid unnatural motion, especially in cases where the existing camera motion is of aesthetic importance.

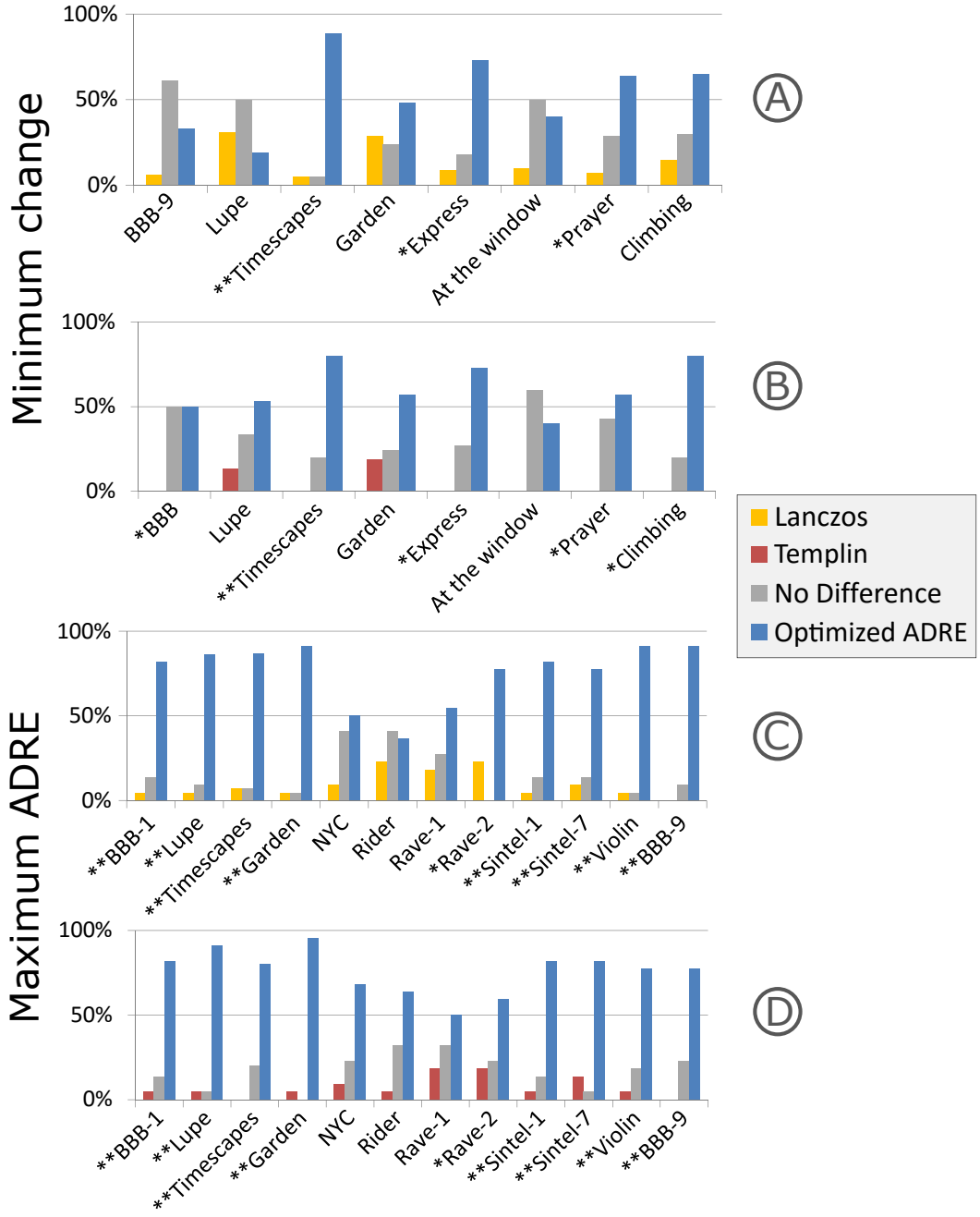

Fig. 4.9 Perceived detail comparison. The participants compared the new approach to Lanczos [Duc79] and Templin [TDR+11] when watching the videos in a loop, and were asked to rate which method was best. Significant differences (χ^2 test) are marked with one or two asterisks, according to a significance level of 95% or 99%. The Lanczos filter is tested with lobe parameters 4, 5 and 6. Cases with the respectively highest rating for Lanczos are listed. Videos were either optimized for minimum change (A,B) or for highest ADRE (C,D).

4.9 Discussion

Based on the findings from the conducted perceptual study on richness of detail, one can infer that, in general, the proposed method is able to achieve a statistically relevant improvement of video quality over the other tested methods considered in this study. The smoothness factor also plays an important role in the perceived quality. Abrupt changes in the trajectory, and especially possible jitter, are strong perceptual artifacts.

Interestingly, most participants did not find out about the synthetically added trajectory if not told beforehand. In fact, the new trajectory resembles the movement of a Steadicam™or hand-held camera which has become increasingly popular for film shooting.[1] Free camera motions filmed with Steadicams™ have become an important visual style in many movies to which the proposed approach can be directly applied and works best because the added motion is not noticed.

In the current implementation, the artificial trajectory applied to the input video results in lost image areas as the video is cropped to the original viewport. Still, in the test scenes with 4K resolution the amount of lost area always stayed below 4.7% of the full frame. The visible area is always above the "action-safe" area of 95% and significantly above the "title-safe" area of 90%. For Full HD videos the value is higher (up to 22.6%) to attain the optimal velocity for ADRE. However, the artist can directly control the amount of lost area for each part of the video by manipulating the importance map and adjusting the time-dependent weights of the importance term. Since undefined areas at the frame borders are seldomly in the focus of the viewer, they are unlikely to attract attention, especially when shown on large screens. Therefore, simple temporal inpainting techniques should be sufficient [BSCB00]. Furthermore, framing is common practice whenever video footage is shown on displays with different aspect ratios. The cropping window can be optimized as the trajectory optimization algorithm automatically provides preservation of the salient region, smoothes the movement of the window, and pays attention to an optimal velocity for ADRE. The cropped area from framing can be used to fill in blank spaces that arise from the approach so that inpainting can be avoided. This problem, however, diminishes if a slightly larger field-of-view is chosen during capturing of a scene or rendering of an animation.

The motion magnitude in the proposed approach is dependent on the video content as well as the resolution of the original video. In most of the performed tests, the high-resolution image was either of size 4096×2304 pixels (4K) or 1920×1080 pixels (Full HD). An increased resolution naturally requires less added motion to the video and hardly changes the intended overall composition of the video. As the market share of home cinema TV panels and projectors in cinemas with Full HD resolution or lower is still over 75%, the proposed ADRE technique could be a valuable solution to increase perceived resolution [Por15].

One current limitation is the inability to faithfully enhance resolution for videos including large semi-transparent objects. The reason for this lies in the inability of current optical flow techniques to

[1] The number of movies using hand-held cameras, or which are shot in the style of "found footage", increased dramatically in the recent two decades. (*cf.* http://foundfootagecritic.com/found-footage-films-database/)

distinguish between different layers. It might be claimed that the underlying assumption of only one salient region in the image restricts the applicability of the approach to only a subset of the possible shots in movies. However, one should keep in mind that fine details in movies, like skin pores or hair strands, are usually only visible in close-up views where the assumption of only one salient region is valid in most of the cases. In the case of a strong salient region the method works well. In contrast to previous techniques the proposed method enables resolution enhancement also for very fast in-video motion limited only by the amount of motion blur recorded in the original video.

4.10 Conclusion

The gap in resolution between capturing and display devices requires downsampling of the original footage resulting in loss of fine detail. The gaze-contingent downsampling approach introduced in this chapter provides an important step towards the preservation of these fine structures. By moving the image along a smooth synthetic trajectory in combination with a temporal upsampling scheme, any video can be optimized for apparent resolution enhancement. Benefits as well as limitations of the approach have been evaluated in perceptual user studies that show that apparent resolution enhancement is achieved even for difficult scenes where previous approaches fail.

There is a number of avenues for future research extending the presented proof of concept. The biggest challenge to bring the ADRE technique to market maturity is to create an appropriate compression scheme of the subsampled videos. A thirty-second Full HD video at 120 Hz has an uncompressed size of more than twenty gigabytes. Encoding the subframes directly is problematic as subframes generated by ADRE exhibit a lot of high-frequency details. Unfortunately, established video compression techniques rely on the assumption that large parts of the video can be predicted by displacing previous and future frames which does not hold for the current algorithm. A promising direction for saving bandwidth would be to compute the subframes in real-time. The overhead for the described technique would be minimal as only the 2D trajectory, two floating point values per frame, needs to be saved in addition to the video. Additionally, super-resolution techniques could be applied to low-resolution videos in order to save bandwidth, as described in [BF12a].

As stated in Section 4.9, ADRE for scenes containing semi-transparent objects is difficult as current optical flow algorithms, in general, assume opaque objects. To support such scenes, a separation into different layers using matting algorithms and tracking of the separate layers is required.

Extending the proposed method to multiple saliency regions with conflicting flows could be possible by treating each region separately and deforming the rest of the image. If the warping is small enough it should not attract attention. Such a deformation was already successfully used for movie reshaping [JTST10].

Although the method concentrates on enhancing the salient regions in movies, it is possible to also enhance manually specified parts of the video by adjusting the importance map. This could be interesting for sports events, e.g., to sharpen some advertisements in the background.

Chapter 5

Perceptual Video Filtering

Contents

When observing a scene, we tend to track moving features to resolve details. Consequently, this tracking leads to perceived blur of the non-tracked objects. A similar observation holds for cameras which record motion blur based on exposure time. The recorded motion blur does not necessarily coincide with the expected perceived blur when watching a video. Especially on displays with a wide field of view, additional eye movement influences perception and can result in visible artifacts such as ghosting, aliasing, and a significant loss of detail. This chapter describes *perceptual blur*, a novel video filtering approach for consistent motion blur computation. The method can also be used to simulate different shutter effects, or for other artistic purposes. It handles real and artificial video input, is easy to compute and has low additional cost for rendered content. A perceptual study using eye tracking demonstrates the advantages and applicability of the method.

5.1 Introduction

Composition, motion, aperture, focus, gain, and exposure time are well-known parameters to artistically influence video recordings [Pro08]. Especially exposure time is an important element as it is inherently related to frame rate. Short exposures lead to discontinuous motion (strobing effect), while longer exposures create motion blur, resulting in detail loss [OS95]. Blur, also in the context of depth of field, can be of significant importance for artistic purposes, for example, to attract attention [Pro08, p. 299], [Bro02, p. 51], to increase motion perception and liveliness [OS95, p. 129], or to serve in story telling [Bro02, p. 62].

It must be realized that perceived blur in the real world will always differ from camera-recorded blur. One of the reasons is that we as humans tend to track the interesting elements in the scene whereas a video camera may or may not follow the same object. Consequently, eye motion and camera motion differ, and so does the corresponding motion blur (Fig. 5.1). Especially for larger screens and low frame rates, this mismatch can result in visible artifacts.

Hold-type blur (*cf.* Chap. 4.2) might occur due to a mismatch between continuous eye movement when tracking an object on the screen, and discontinuous movement of the target due to limited frame rate. The latter can be very confusing as the human visual system (HVS) expects tracked objects to move smoothly and to appear sharper than non-tracked objects.

Current high frame-rate (HFR) videos, with a typical frame rate of 48 Hz to 60 Hz, reduce recorded motion blur and hold-type blur, leading to sharper perceived images. For this reason, they have become popular in the consumer market; specialized upsampling techniques are integrated into standard TV equipment, and high frame-rate movies are being explored by movie directors (e.g., *The Hobbit*). Nevertheless, the consequences are not always beneficial. An HFR video must be recorded at lower exposure times and, because there has to be a minimal time to store a frame (or to open the shutter), the shutter is open only 60% of the overall time [Bro02]. Temporal replicates caused by high sampling rate, and perceivable as shifted ghost images, may appear. Some viewers even report perceiving a distracting speedup of the video [Fen14]. For this reason, recent HFR television shows

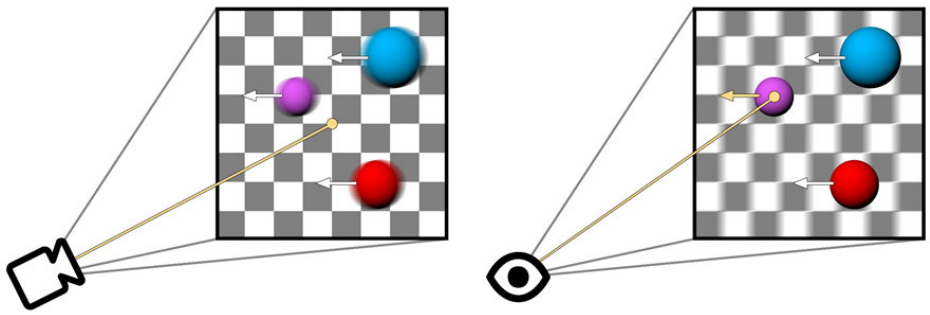

Fig. 5.1 Blur mismatch. In case camera motion (left) and eye motion (right) differ due to smooth pursuit eye motion, the mismatch in corresponding motion blur can result in visible artifacts.

such as *Video Game High School* added hand-tuned blur to some scenes, thereby removing many of the details. Such solutions are rather ad-hoc and not always successful.

Whether considering motion or hand-created blur, the blur does not lead to the expected perceptual blur induced by eye movement. Even a frame rate of 120 Hz—far beyond the 48 Hz used in current HFR movies—is insufficient to allow for natural perception of blur, not to mention the high bandwidth requirements and lack of support by current displays [Tru13]. To remove the camera-induced motion blur, an infinitely high frame rate would be required. Hence, arguably the only practical solution is to include the respective eye motion into the blur model and to create a displayable lower frame rate video from an ultra-high frame rate video (UHFR) based on the expected eye motion. [1]

To solve these problems, in this chapter a novel method is presented to adapt exposure and motion blur in a postprocess by taking eye motion into account. A perceptual model is derived to explain the perception of a scene from a standard video camera (Section 5.3). As it is shown in Section 5.4, the camera itself is an insufficient approximation of human perception when content is tracked in the image plane by the observer. As a solution to this issue, a filtering technique is derived that takes predicted eye motion into account, and leads to a more faithful image reconstruction on the retina (Section 5.5). The solution can be used to artistically manipulate frame rate and exposure time in a postprocess, which goes beyond the possibilities of a standard camera.

Specifically, the following contributions are provided:

- a model for perceived motion blur
- a gaze estimation algorithm and corresponding filtering process to create a more faithful retinal image

The perceptual blur provides a variety of benefits and applications, including downsampling for real-world and CG-generated UHFR videos, virtual shutter simulation, motion stills generation, and subtle gaze direction. The technique is applicable to high-speed footage as well as traditional LFR camera output (24–30 Hz) or synthetic content and leads to improvements in perceived video quality

[1] The term UHFR is used to depict videos with a frame rate higher than 1000 Hz.

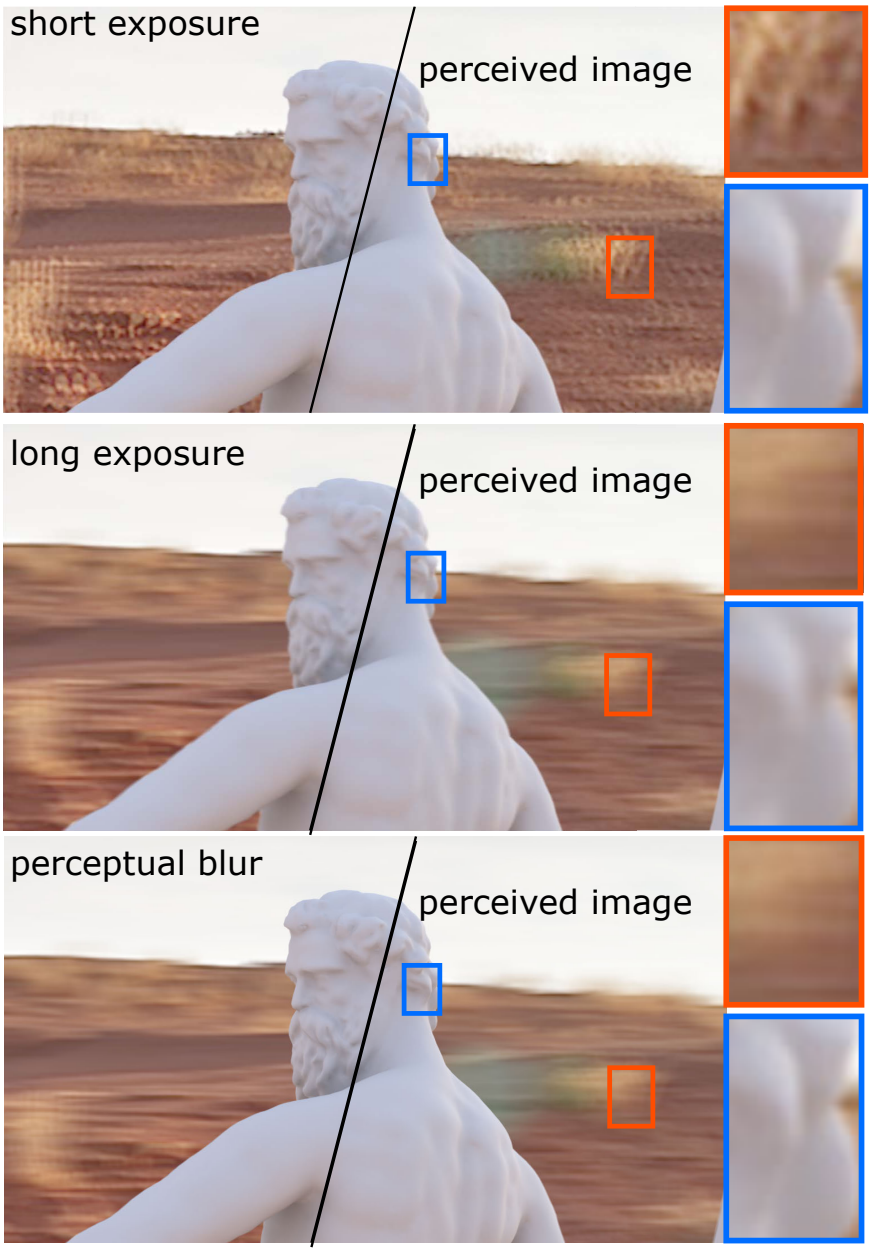

Fig. 5.2 Exposure comparison. Ghosting artifacts are perceived if exposure time or frame rate are insufficient (top row). Longer exposure times (second row) avoid ghosting, but details in the scene suffer due to motion blur. Further, this blur does not match the expected motion blur of an observer watching the scene. Perceptual blur (third row) is a temporal downsampling method that takes eye motion into account. It leads to sharp tracked objects and consistent motion blur in the rest of the scene. The method also makes it possible to subtly guide gaze.

(Section 5.6). For rendered scenes, the solution does not necessarily require higher computation times. The benefit of perceptual blur for subtle gaze steering is illustrated in a user study.

5.2 Temporal Video Filtering

Shutter and exposure time are usually set during the capture process. In contrast, the approach described here modifies these parameters in a postprocess. Therefore, the method can be seen as a generalization and extension of synthetic shutter speed [TSY+07] which imitates a long exposure shot by taking a series of short exposure photographs and aligning them, e.g. to reduce noise and camera shake while preserving motion blur. By taking the viewer into account, the novel filtering also differs from traditional temporal downsampling of HFR videos [FCW+10].

Inspiration is drawn from rendering techniques to simulate shutter effects [Gla99]. Previously, the desired shutter type had to be chosen beforehand so that any change implied a costly reshooting of the scene. From an artistic perspective, a postprocessing solution is much more desirable and makes it possible to freely test different shutter types to find the desired final appearance.

A rolling shutter determines exposure time via an opening of a rotating disc. An open arc of 180° has been established for 24 Hz shots [OS95]. Shorter shutters create stuttered motion which may be used for artistic purposes (e.g. 45° in "Saving Private Ryan", "Gladiator", "Three Kings"). A longer exposure (210°), especially in combination with low frame rates (6–12 Hz), creates dramatic blur effects.

Blur can also help in guiding the observer's gaze. Our foveal vision features a high density of cones and leads to high acuity compared to peripheral vision [Ost35]. The latter is still sensitive to subtle temporal changes and can attract the viewer's attention [BMSG09, MBG08], which is another motivation to avoid temporal artifacts. The HVS is attracted mostly to salient regions (*cf.* Chapter 3.1.2). Special blur and sharpening filters [Mit04], depth-of-field effects [KMH01], or observer-driven simplification [DS02b] are good means to accentuate or subdue saliency and to guide gaze. The proposed technique allows to add such indications.

Temporal processing influences our blur perception; strongly blurred ($>$10 arcmin) moving patterns on a tracked objects (SPEM) appear sharper than their static counterpart, yet for a small blur ($<$10 arcmin) stationary edges seem sharper than moving ones [HGG98]. This phenomena is known as *motion sharpening*. It is not a mechanism that removes blur but results from the HVS's inability to discriminate whether or not the moving object is indeed sharp [BM97]. This theory has been strengthened by the fact that observers tend to match blurred peripheral stimuli with sharper foveal stimuli [GOSG97]. Temporal and spatial coherence, as well as motion contrast, are important factors for the HVS also in video saliency [LZDY08, ZS06]. Thus, it is proposed to blur the video according to the predicted/intended eye motion instead of camera motion.

Although eye-movements are not known a priori. Dorr et al. report that in natural movies up to 80% of the subjects look at the same image region [DMGB10]. Especially in Hollywood movies, coherence was very high due to camera work and scene cuts, and target regions are of high saliency [BDK+06, BKBM04]. As the perceptual blur may reduce saliency in the areas outside the object of interest, it can be seen also as an extension to classic movie techniques used to subtly draw attention to specific regions.

In movie productions, the widest rolling shutter is limited to 210° due to minimum sensor read-out time. Consequently, each recording shows gaps that can result in motion artifacts such as judder (unsmooth motion) or edge banding (overlapping edge replicas at the borders of a moving object). It is possible to analyze the required sampling rate and maximal motion between two images to prevent these replicas [CTCS00]. Hoffman et al. propose employing a multi-flash protocol for videos to reduce artifact visibility at a given capture rate [HKB11]. This idea is adopted when constructing an initial ultra-high frame rate video, free from temporal artifacts. Another option is to employ an appropriate bandlimiting filter [SYGM03]. Templin et al. propose a video resampling approach, reducing perceptual video artifacts [TDMS16]. The authors make use of an automatically computed "frame rate map" used for filtering the video in order to emulate arbitrary continuous frame rates for videos. To remedy the effects of temporal aliasing, Nvidia recently released its G-Sync technique which adapts display refresh rate to the processing time of each rendered rame[2]. The technique reduces some stutter motion artifacts by avoiding repeated display of the same frame allowing for smoother motion perception in games.

5.3 Image Formation Model

Even though many directors consider the video camera as the observing eye, watching the video afterwards does not create a perfect illusion of viewing the recorded scene as in the real world. In the following a model is described to predict how a real scene and its captured video are perceived by the human visual system (HVS).

For ease of explanation, the irradiance that is recorded on the sensor plane of a video camera is defined as a function $\mathbf{S}(x,t)$, where x is sensor position and t is time. Ideally, the recorded video $\mathbf{I}$ would be equal to $\mathbf{S}$. However, it is a discretized version. We focus on temporal discretization and assume resolution to be high.A frame $\mathbf{I}_i$ is described as

$$\mathbf{I}_i(x) := \int_{t_i}^{t_i+T_V} \mathbf{S}(x,t)dt \quad , \tag{5.1}$$

where the camera shutter opens at time t_i and closes again at time $t_i + T_V$. A shorter open shutter reduces intensity, but it is assumed that gain is used to counterbalance exposure variations. Based on the common usage of a 180° shutter in traditional film-making, the exposure time T_V is chosen in accordance with the simple equation $T_V = 1/(2 * \text{frame rate})$.

Analogously, the retinal image $\mathbf{R}$ is defined by the intensity perceived on the retina. We define $\mathbf{R}$ for a retinal location x as

$$\mathbf{R}_i(x) := \int_{t_i}^{t_i+T_R} \mathbf{S}(x+p(t),t)dt \quad , \tag{5.2}$$

where $p(t)$ describes the eye's path due to SPEM tracking, and T_R is a small period of time over which the information is integrated by the HVS. T_R is referred to as the *critical duration* which is inversely related to the critical flicker frequency (Chapter 2.3). The critical duration is an empirically estimated value which can be imagined as how long a receptor of the retina accumulates incoming

[2] http://www.geforce.com/hardware/technology/g-sync

photons before an electrical stimulus is triggered for higher-level processing in the retinal ganglion cells [AKLA11, p.699]. The critical duration depends on incoming light intensity. This dependency is described by Bloch's Law [Blo85] (Chapter 2.3). Since for photopic vision rods are fully saturated, the perceived signal only depends on the cones. Considering these aspects in Eq. (5.2), it is assumed that the critical duration T_R to be 15 ms, which is the longest temporal summation time for cones [Bur81]. p is closely linked to feature tracking—saccades and tremors can be neglected because *smooth pursuit eye motion* allows us to almost perfectly track targets up to object speeds of about 10 deg/s. Higher speeds may lead to significant differences in perception between individuals [RG09]. The combination of smooth pursuit eye motion and the integration time of the HVS explains hold-type blur [PFD05] which results from the mismatch between (discontinuous) object motion on the screen and (continuous) eye tracking. It is particularly pronounced for low frame rates. More details on capture and display of movies in signal processing terms are provided by Watson et al. [Wat13].

Photoreceptors do not move independently but are fixed on the rigid retina [CSKH90]. Hence, it can be safely assumed that p is the same for all locations on the retina. While this is not exactly true (the induced error depends on eye shape, viewing conditions, and camera settings), any deviations are negligible for the purposes of the proposed approach.

5.4 Blur Mismatch of Camera and Eye

A typical low frame-rate video camera is an imperfect substitute for the human eye when an object of interest (OOI) moves in the image plane. The reason is that the additional eye movement while watching the video should have been taken into account during recording. Let the exposure time be equal to the integration time T_R of the HVS which is reasonable for 30–60 Hz videos [Blo85]. Then, only in the absence of any eye movement, Eq. (5.2) is equivalent to Eq. (5.1), i.e., $p(t) = 0$.

A simple example is shown in Fig. 5.2. Camera motion is an off-axis rotation around the OOI resulting in both a rotation and translation of the Neptune statue in image-space. For short exposure times, the OOI is detailed but the background exhibits temporal artifacts (left). These artifacts appear only on the retina of the observer; they reveal themselves as an unnaturally sharp background or even ghosting whenever the integration time of the eye crosses frame boundaries. For long exposure times, the OOI suffers from motion blur (middle).

To explain the temporal artifacts, we consider a simple scene (Fig. 5.3); a small object moves horizontally from left to right while the eye tracks an OOI that moves vertically in the image plane. The correct integration on the retina should result in a diagonal line (Fig. 5.3b). If the camera is static and captures at a frame rate equal to $1/2\ T_R$, the object is smeared along a horizontal line due to the recorded motion blur (Fig. 5.3c). When watching the video, however, the eye tracks the upward moving OOI. This eye movement results in hold-type blur. The retina integrates at each location the pixels that are crossed, resulting in two square-shaped features on the retina, one for each frame (Fig. 5.3d).

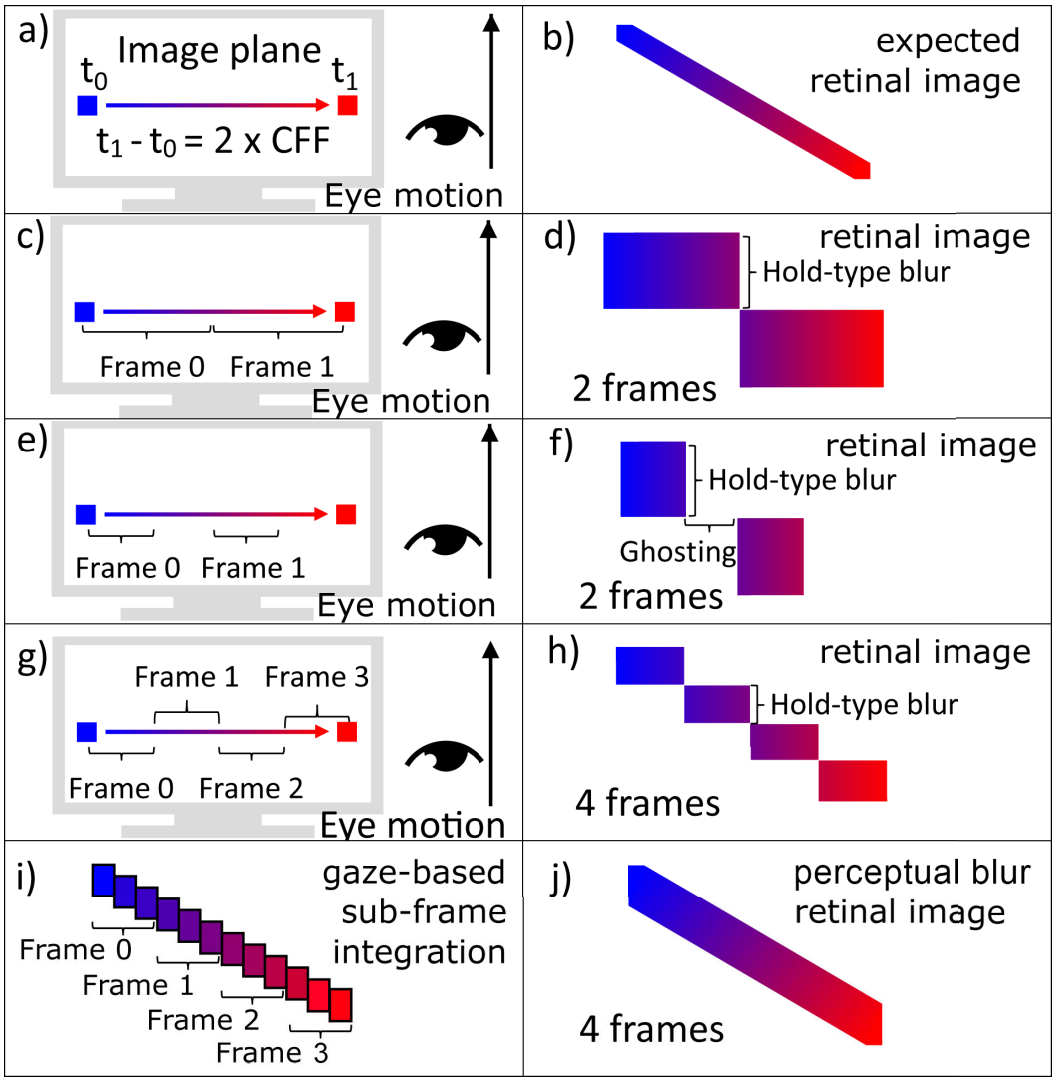

Fig. 5.3 Temporal artifacts. (a) A small object moves horizontally, the eye tracks upwards. (b) Expected perceived image. (c) Long-exposure recordings with a static camera transform the point in a horizontal streak. (d) However, due to eye integration, an observer perceives two rectangles which are only loosely connected, although they come from the same object. (e) Shorter exposure times lead to temporal artifacts; separate components are perceived and temporal information is lost. (f) Two disconnected rectangles appear on the retina. (g) High frame-rate videos exhibit more frames and less motion blur, which can reduce the problem. (h) By increasing the frame rate, one can approximate the expected retinal image, but the needed frame rates are not possible to capture without time gaps, and displaying them is challenging. The proposed perceptual blur makes use of the eye motion and results in a closer approximation to the expected retinal image (b) for (i) low as well as (j) high frame rates.

To reduce motion blur for important objects, a shorter exposure time can be used, but then some temporal information is lost (Fig. 5.3e). This leads to perceivable ghosting, and features seem to *jump* (Fig. 5.3f). Only by increasing the frame rate (Fig. 5.3h) hold-type blur is reduced linearly and, in the limit, converges to zero. Nonetheless, it is difficult to record videos at such high frame rates without gaps. Also, displaying the content is challenging due to the necessary high bandwidth. Our temporal downsampling, explained in the next section, simulates perceptual blur of the HVS and delivers a more faithful image on the retina (Fig. 5.3i,Fig. 5.3j).

5.5 Gaze-guided Downsampling

Temporal edge banding or ghosting artifacts in videos only appears if the projected scene motion between two frames exceeds one pixel, a result that can be derived from similar findings for light field rendering [CTCS00]. Hence, although the goal of the method is consistent filtering and temporal downsampling, if LFR or HFR video footage is given as input, the video sequence is first transformed into an utra-high frame rate (UHFR) video $\mathbf{I}_{\text{UHFR}}$ with frame rates of 1000+ Hz. This footage is computed by using an interpolation algorithm based on image similarity [LLN+10]. This upsampling process is usually robust but can fail for blurry edges. Fortunately, blurry edges are not very problematic for the eye integration process [DER+10a] and are unlikely to exhibit temporal edge banding. Nonetheless, to achieve sharp images, the original exposure time should be short.

If the eye motion p is known or user-defined, the retinal image $\mathbf{R}$ can be computed for any point in time and any desired integration time T_R by integrating $\mathbf{I}_{\text{UHFR}}$ along p. More intuitively, this is equal to translating each frame along the opposite direction of p. Basically, this compensates for eye motion and sets the motion difference to zero. Next, integrating along the temporal axis and translating the frames back to their original position results in the filtered video frames.

A method for perceptionally plausible reconstruction while temporally downsampling $\mathbf{I}_{\text{UHFR}}$ has been introduced for apparent display resolution enhancement (Chapter 4.4): the retinal image is computed and the output video frames are optimized such that the integration in the eye best matches the target image. Nevertheless, because the output video is not necessarily an UHFR sequence, exploiting eye integration is difficult and frequencies that would cancel out any hold-type blur may not be created .

Instead, a frame $\mathbf{R}_i$ of the output video sequence is derived via

$$\mathbf{R}_i(x) := \int_{t_i}^{t_i+T} \mathbf{I}_{\text{UHFR}}(x + p(t), t)dt \tag{5.3}$$

where T is the desired output frame duration (inversely related to the frame rate). Hereby, an image similar to the expected retinal image is produced. As p is defined via the OOI, this part of the output stays sharp. Further, when assuming hold-type blur, the result stays consistent; non-tracked objects will be consistently blurrier than the OOI. The definition also leads to robustness with respect to eye tracking because any deviation from the intended path will not introduce additional high frequencies, as could be the case in [DER+10a]. In Section 5.6, it is shown that these properties help to avoid

temporal artifacts and to improve perceived quality. Further, the effect is useful for guiding the gaze of observers.

Saliency-based Temporal Integration

Eq. (5.3) requires the eye's motion path p to be known, which is the case if it was intentionally created and imposed for artistic purposes. In all other cases, p must be predicted by saliency.

A saliency map $\mathbf{A}$ is computed for each frame of $\mathbf{I}_{\text{UHFR}}$ [HKP07], with $\mathbf{A}$ being normalized to the range $[0,1]$ which gives a direct measure of how probable it is for a viewer to track a feature in $\mathbf{I}_{\text{UHFR}}$. The gaze path $p(x,t)$ is assumed to correlate to the optical flow $\mathbf{F}_{i\to i+1}(x)$ [LLN+10, ZPB07] for any two frames i and $i+1$ in $\mathbf{I}_{\text{UHFR}}$, which proved successful in Chapter 4.5 as well. Nonetheless, it is not a robust solution to simply select the path $p(x)$ of the most salient x to define a global p. Instead, a set of pixels describing the OOI is marked. The mask is derived from the likelihood of each pixel to belong to the OOI. Precisely, the algorithm estimates the region of high saliency and similar pixel motion [HKP07]. For frame i, the mask $\mathbf{O}_i$ is defined as:

$$\mathbf{O}_i(x) := \begin{cases} 1, & \text{if } ||\mathbf{F}_{i\to i+1}(x) - \frac{\sum_x \mathbf{A}_i(x)\mathbf{F}_{i\to i+1}(x)}{\sum_x \mathbf{A}_i(x)}|| < \tau \\ 0, & \text{else} \end{cases} \tag{5.4}$$

In practice, a value of $\tau = 7$ works well and was used for all examples. The gaze path $p(t_i)$ is set to the movement of the center of masses from $\mathbf{O}_i$ to $\mathbf{O}_{i+1}$. Additionally, manual restriction of the OOI is allowed by creating an optional binary mask $\mathbf{M}$ that is transferred from one frame to the next via rotoscoping [AHSS04]. Hereby, multiple OOIs can be disambiguated when needed which is also useful for artistic purposes and allows choosing and guiding gaze direction (Section 5.6.6).

Although only translational motion is considered for eye integration, this choice is not very restricting. The translations are applied to an $\mathbf{I}_{\text{UHFR}}$ sequence. Hence, each frame exhibits minimal motion. Furthermore, the sequence itself was constructed via an upsampling technique that assumes general motion. Although it is true that different parts of the OOI can undergo different motion, ultimately, our eyes can only follow a single path which for a single OOI is usually well detected by the described method [DMGB10].

5.6 Applications

This section shows results and applications for the proposed gaze-contingent temporal resampling method. Precisely, results of the perceptual blur are compared to video sequences captured with *short exposure* times (pinpoint-sharp images with the typical 180° shutter, resulting in an exposure time of half the frame duration) and to *long exposure* shots where the shutter was kept open for as long as possible. The shown video footage includes synthetic as well as real-world sequences captured with traditional low frame-rate and high-speed cameras.

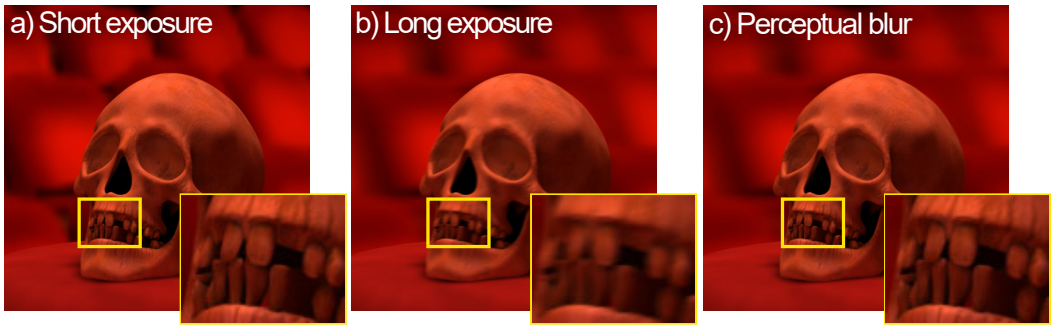

Fig. 5.4 Synthetic Ultra-high Frame-Rate Video: (a) short exposure, (b) long exposure, (c) perceptual blur using a rendered video (60Hz).

Fig. 5.5 Real-World Ultra-high Frame-Rate Videos: short exposure (left column), long exposure (middle column), perceptual blur (60 Hz, right column) for ultra-high frame rate input (3000Hz).

Fig. 5.6 Low Frame-Rate Video: The 60 Hz video was upsampled to 3000 Hz, then downsampled to 60 Hz to simulate different exposure times. (a) Original. (b) Long exposure; entire image is blurred. (c) Perceptual blur; OOI is kept sharp while background is blurred.

5.6.1 Ultra-high Frame-Rate Videos

First, the downsampling results are illustrated for two synthetic "hero" shots (Fig. 5.2,5.4): the camera moves around the object of interest in an ellipse, creating opposing foreground and background motion. Short exposure leads to temporal artifacts in the background which results in ghosting artifacts on the retina when tracking the foreground. A long exposure shot removes these artifacts but blurs the foreground. Perceptual blur leads to sharp foreground while avoiding background ghosting. A static camera was explicitly used to show the difference between a short/long exposure shot and the perceptual blur. The short exposure shot keeps both fore- and background in focus which results in unnaturally sharp images. The long exposure, on the other hand, removes most details from the foreground object. Gaze-contingent downsampling filters the background slightly to counteract the hold-type blur and puts emphasis on the main elements.

5.6.2 Stochastic Ultra-high Frame-Rate Videos

For CG-generated sequences, the upsampling process can be modified and an UHFR sequence can be derived more easily by relying on existing temporal coherence methods [SYM+11]. In particular, given a physically-based renderer, which have become common in production rendering [Sol16, ENSB13], a low quality UHFR video is created as input to the temporal filtering algorithm (Fig. 5.7b). Each frame of the low-quality video is rendered with just a fraction of the samples required for a high-quality solution, thus not changing the overall number of samples required for rendering the LFR. The proposed filter kernel gathers samples over multiple frames of the UHFR video, resulting in a high-quality LFR video with the desired motion blur (Fig. 5.7c). The validation of this step is related to distributed ray tracing [CPC84]. It also implies that for physically impossible exposure times that exceed the duration of a frame, compute time decreases using the proposed filter (Fig. 5.7d). This behavior is different from most Monte Carlo-based motion-blur rendering techniques where stronger motion blur tends to increase render times [NSG11].

5.6.3 Low Frame-Rate Real-World Videos

Low frame-rate videos are first upsampled to UHFR by relying on a standard temporal upsampling technique (Sec. 5.5), then downsampling is applied (Fig. 5.6). If the target frame rate is equal to the original frame rate, the gaze-aware filtering algorithm uses the original frames inside the OOI and the interpolated frames in the background. This strategy avoids artifacts in the OOI induced by potentially imperfect upsampling, e.g., upsampling may fail to produce faithful results in case of dis-/occlusions.

5.6.4 Virtual Shutter

The presented temporal filtering approach is compatible with virtual shutter simulations. Rolling shutter, focal plane shutter, or even artistic shutters can be simulated. On a per-pixel basis, the exposure interval of Eq. (5.3) is defined, resulting in a direct integration into the described solution.

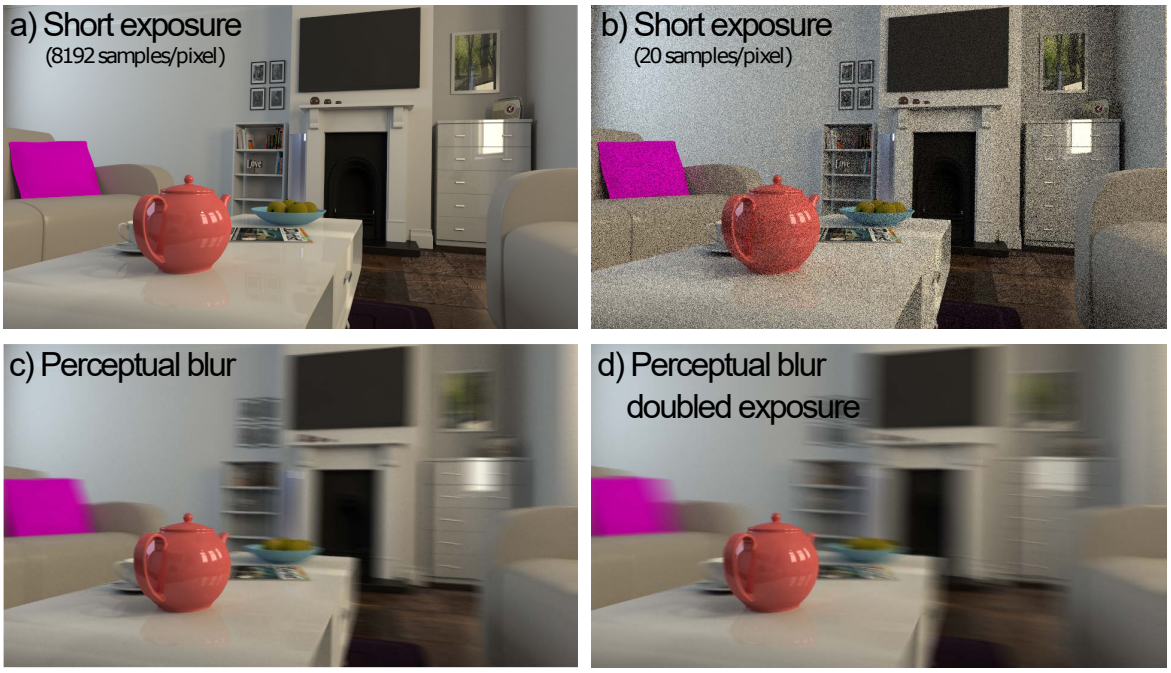

Fig. 5.7 Stochastic Ultra-high Frame-Rate Videos: (a) High-quality short exposure (8192 samples per pixel). (b) Image from low-quality, high frame rate video (20 samples per pixel). (c) Applying the gaze-aware downsampling to (b) leads to approximately similar quality as in (a). (d) Physically impossible exposure (twice the frame time) using only 10 samples per pixel as input (both 30Hz).

In Fig. 5.8, a focal plane shutter was used to imply speed by producing a tilting effect. In this case, the per-pixel definition was given by shifting the time interval in each row (top to bottom).

5.6.5 Motion Stills

Images represent a snapshot in time. However, a single time slice or pinpoint sharp image does not convey any information about the motion in the scene. In contrast to short or long exposed traditional imagery, perceptual blur keeps the OOI in focus but still preserves important motion blur information (Fig. 5.2–5.8). This is especially interesting for advertisements or movie descriptions in magazines where one wants to convey the dynamics of the scene.

5.6.6 Subtle Gaze Direction

To investigate the influence of perceptual blur on gaze behavior of an observer, a perceptual study was conducted. As stimuli, identical spheres are shown which move at equal speed in different directions, Fig. 5.9a,b. This simple artificial scene was intentionally chosen to reduce the influence of higher-level perception mechanisms as much as possible. Each video is twelve seconds long, created via our downsampling method from a 3000 Hz input video to 60 Hz, but focusing on different spheres as OOI

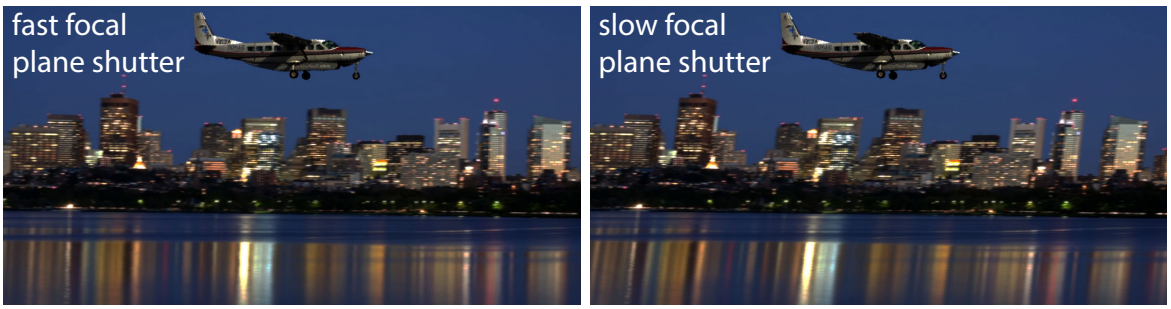

Fig. 5.8 Shutter postprocessing The temporal filtering approach allows to redefine shutter types after recording. Here, a focal plane shutter with different speeds is applied to a synthetic scene.

(Fig. 5.9b). Two sets of videos have been created, one with $1/60$ s and one with $1/30$ s exposure time. For both, an additional long exposure shot was created as a reference (Fig. 5.9a) to validate whether the filter has any measurable influence on eye-motion behavior. Additionally, the stochastic-rendering ROOM scene, Fig. 5.7, was shown, once with a $1/30$ s traditional long exposure time and once with the gaze-aware downsampled version using the same exposure time but focusing on the teapot (Fig. 5.7d). Each sequence is about one second long (76 frames, 60 frames per second). These eight videos were shown three times to each participant in randomized order.

14 participants, unaware of the goal of the experiment and with normal or corrected-to-normal vision, took part in the experiment. As display Samsung RZ2233 Full-HD screen was used in a darkened room to present the video footage, and an EyeLink 1000 eye tracker was used to record fixation times on the screen. The eye tracker recorded at 1000 Hz. The participants were seated in 60 cm distance to the screen. The participants did not receive any specific task except for watching the videos to prevent any task-specific influence of the results. The test took around ten minutes for each participant.

Synthetic Scene Fig. 5.9 shows snapshots and results of the experiment in form of gaze heat maps. The top row shows the long exposure shot (left) and one version using perceptual downsampling (right) where the focus was on the diagonally moving ball marked on the left. Below are heat maps describing the average gaze distribution for all participants, again for the long exposure shot and the novel approach with $1/30$ s exposure settings. Perceptual downsampling to $1/30$ s increases fixation times for the intended objects of interest (Fig. 5.9d). A statistical evaluation (Fig. 5.10a) reveals that the fixation time is roughly even among all spheres in the long exposure video, with a bias towards the central objects.For a simulated standard camera with $1/30$ s exposure, participants followed the selected OOI for 14% of the time with focus on the diagonally-moving sphere near the left border (sphere A, standard deviation SD=7.2%, increase for 11 out of 14 participants) and 26% with focus on the horizontally-moving sphere (sphere B, SD=9.0%, increase for 11 out of 14 participants). Using the gaze-contingent approach, the average percentage increased to 32% for sphere A (two-tailed t-test p=0.0015) and to 44% for sphere B (p=0.0013) which is a significant relative increase by 124%

and 68%, respectively. These results strongly indicate the ability of perceptual blur to influence gaze. For an exposure time of 1/60 s, the effect is more subtle, changing from 12% (*SD*=5.9%) to 16% (*SD*=6.2%, sphere A, increase for 10 out of 14 participants, two-tailed *t*-test p=0.12) and 21% (*SD*=7.6%) to 29% (*SD*=10.8%, sphere B, increase for 12 out of 14 participants, p=0.04) which is still a relative increase by 31% and 36%, respectively.

Complex Scene In the more realistic ROOM scene (Fig. 5.7) the camera rotates around a view-centerd object (fruit bowl) while the perceptual filter is applied to focus on an off-centered object (teapot). Since the camera rotates quickly around the center of the scene, off-center objects appear strongly blurred in the long exposure video. However, the object in the center only suffers from blur caused by rotation and therefore remains sharp. In the study two versions of the scene were shown, one long exposure video and one video filtered by the gaze-contingent method. In the filtered version the teapot was selected as the object of interest. It was hypothesized that the central object (being a fruit bowl, as visible in Fig. 5.7a) would mainly attract the attention of the viewer in the long exposure version.

A statistical evaluation of the fixation time on the OOI is shown in Fig. 5.10b (significant differences marked by $*$). In the very beginning (frames 1 to 5) of both versions of the video, the participants fixate on the center of the screen due to the earlier calibration step. In the following (frames 6-20) the teapot moves into the central part of the screen caused by rotation of the camera, and it starts attracting attention. As the teapot moves off center again from frame 25 on, fixation time on it decreases. In the long exposure video, most participants focus on the sharper center object. The amount of time the subjects fixate on the teapot in frames 21-26 and 27-66 decreases to 28% and 7%, respectively.

In the filtered version the fixation time in frames 21-26 and 27-66 increases significantly to 67% and 25% in total, which is a relative gain of 139% (M_{long}=27.9%, SD_{long}=12.1%, M_{perc}=66.7%, SD_{perc}=15.6%, two-tailed *t*-test p=0.0016) and 223% (M_{long}=7.8%, SD_{long}=5.8%, M_{perc}=25.2%, SD_{perc}=6.1%, p=0.001). Towards the end of the video (frames 67-76) the rotation of the camera slows down and finally stops. Since there is no motion blur without motion, all scene objects appear sharp. The gaze analysis reveals that the participants change their focus to diverse objects in the environment of the scene. Accordingly, fixation times for the long exposure video and the filtered version converge to the same level as there is no visual difference between them. In total, the overall fixation time on the OOI increases for 12 out of 14 participants. It is likely that the temporal sensitivity of the human peripheral vision influences the participants' focus because high frequency video content tends to attract gaze [Bur81]. The results of both perceptual studies suggest that the effectiveness of gaze guidance using the perceptual filter increases with exposure time, even beyond physically possible exposure times, emphasizing the importance of being able to adjust exposure in a post-process.

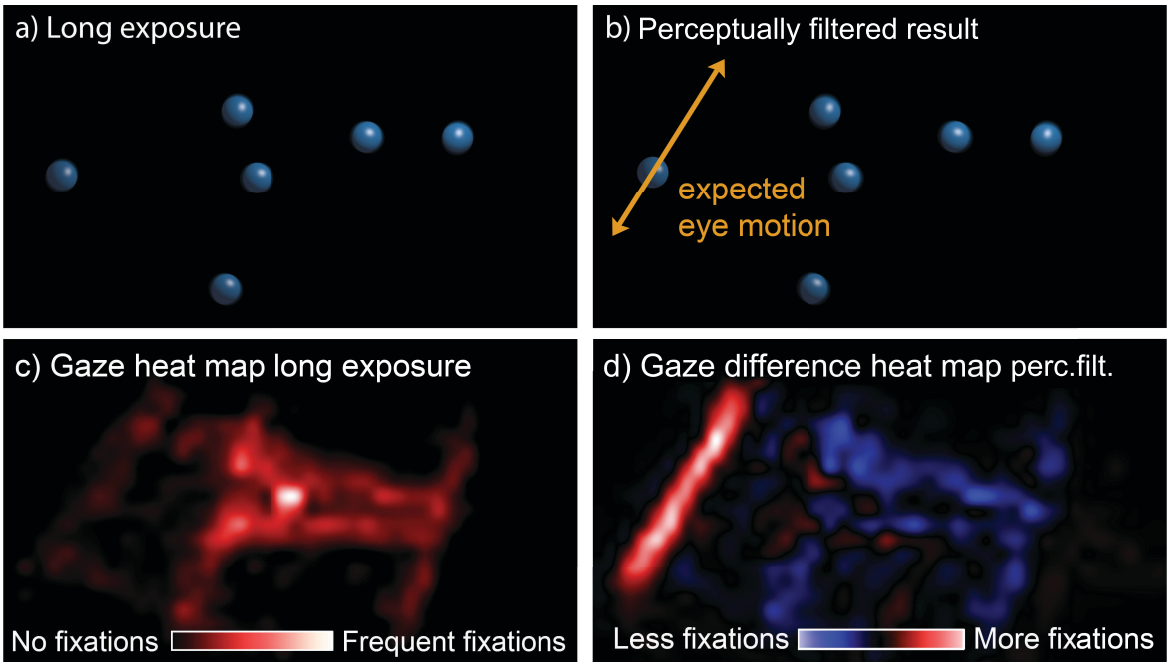

Fig. 5.9 Subtle Gaze Direction: (a) Image from the long-exposure sequence. (b) Image from the perceptually filtered result. The expected eye motion induced by the OOI is shown in orange. (c) Gaze heat map of tracked gaze direction for the long-exposure sequence. (d) Relative gaze difference for the filtered result related to long-exposure sequence. The OOI (sphere A) exhibits an increased fixation time (red areas). Other spheres are fixated less (blue areas).

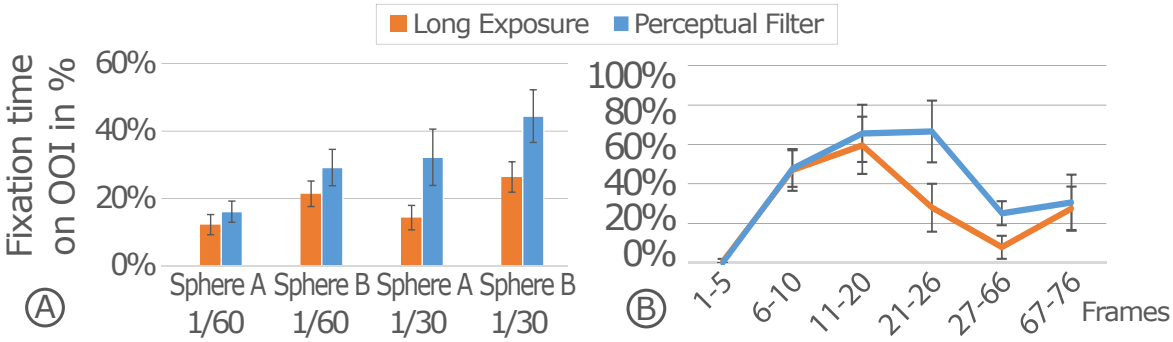

Fig. 5.10 Quantitative evaluation of SPHERES and ROOM sequences: (a) The time the users spend on the specified sphere for the given exposure time is given in percents of the total video length. In the perceptually filtered video (blue) the fixation time is increased compared to the long exposure video (orange). (b) In the respective frame ranges of the ROOM video sequence the colored lines represent the average amount of time the users spend on the specified object of interest in the two version of the video (long exposure, orange; perceptually filtered, blue).

5.7 Discussion

As indicated by the results and perceptual study, the proposed method makes artistic postprocessing possible and can successfully influence observers' gaze. Humans rather focus on sharp and moving objects when watching videos. Hence, knowledge of a reasonable scanpath is not necessarily required but can be created with the proposed technique. This may develop into an important tool for movie production.

There is a trend towards large-screen home theater systems and wide FOV displays. Since the artifacts induced by traditional cameras are more obvious on larger screens, the difference of a long-exposure and a perceptually-filtered video becomes more pronounced, which renders the described filtering method increasingly interesting.

The required UHFR videos have a non-negligible memory footprint. In most cases creating a full UHFR video can be avoided by using a cache maintaining the necessary UHFR frames and whose size depends on the desired integration time.

The perceptual filter does not yet address hold-type blur which reduces higher frequencies in the direction of eye motion. For lower frame rates, one would need to introduce higher frequencies into the images that would cause visible artifacts in the still images [DER+10a]. Instead, the method computes a consistent image, reducing any temporal edge banding artifacts while keeping the OOI sharp.

The proposed method assumes frame duration times below the integration time of the HVS. A video played at very slow frame rates may result in discontinuous motion on the retina due to insufficient eye integration in this case. Finding a solution for this case is a problem on its own.

The accuracy of the automatic saliency metric works well for the tested scenarios but is not perfect. If it fails, in the worst case, attention may be drawn to different parts of the video. In addition, standard tools for matting and rotoscoping can always be used to correct or manually define saliency masks. Often these have already been created for other post-processing steps, such as color grading or 2D-to-stereo conversion, and may be available in most production settings.

Assuming a single OOI should not be considered a strong limitation because this restriction holds similarly for any standard camera. Further, if selecting multiple OOIs the proposed method computes an average motion for all OOIs creating a video that keeps them in focus as good as possible, at least as good as for the case of a traditional camera.

Another assumption inherent to the OOI is that it will not be occluded by another fast-moving object. Potentially, these situations can lead to conflicts. In practice, these are typically the situations when tracking becomes more difficult for an observer and they tend to be more forgiving with respect to temporal artifacts, as their tracked signal is expected to be discontinuous. Not following the intended scanpath has an effect but it does not necessarily deteriorate the viewing experience. Perceptual blur does not produce a blurrier overall image; objects moving in approximately the same direction as the OOI will appear sharper than with standard long exposure. However, if the viewer deviates from the intended scanpath, temporal flickering can theoretically occur for the OOI - although no participant

reported such observations in the conducted perceptual study. This subtle effect might actually help to guide the gaze towards the OOI, similar to [BMSG09]. It is convenient that as soon as the observer follows the OOI, gaze follows the intended path and potential artifacts will disappear.

Having access to UHFR footage is a benefit because the right tradeoff between exposure time and motion blur is often difficult to decide upon when capturing a scene. Especially for stunt shots, there are many fast movements and repeating the action can be very costly. HFR equipment is currently expensive, but hardware prices drop and movie makers start recognizing the new possibilities and advantages. It is difficult to answer whether a higher frame-rate movie or a perceptually-motivated motion blur "looks better". We are conditioned to Hollywood movies recorded at 24 Hz, and the audience reacted reluctantly to the 48-Hz version of "The Hobbit" as they were not used to the new viewing experience. However, there is a clear tendency towards higher frame rates (e.g. "Avatar 2" by James Cameron will be shot at 60 fps) and it is crucial to investigate this area in depth. The solution described in this chapter is a first significant step in this research field.

5.8 Conclusion

The temporal integration in traditional camera recordings does not correspond to the integration of the human visual system when watching a movie. In this chapter, a gaze-guided as well as gaze-guiding, temporal downsampling technique was proposed to achieve consistent results without edge banding or judder artifacts, for real and synthetic video input of arbitrary frame rate. A model for video perception based on the human visual system was introduced. The proposed gaze-guided downsampling approach uses video saliency. Different applications of perceptual blur have been presented, including downsampling of real-world and CG-generated ultra-high frame-rate videos, virtual shutter simulation, motion stills generation, and subtle gaze direction. A perceptual study confirmed the effectiveness of the approach to influence observers' gaze.

One future direction is to support multiple objects of interest also via an interpolation of the eye motion vectors on the image plane. A Poisson reconstruction using the OOIs as boundary conditions could be an option. Nonetheless, in practice, assuming a single OOI currently leads to better results, and the proposed method is robust with respect to deviating eye motion. It was shown that the approach enables interesting post-processing possibilities. Many more applications could possibly benefit from perceptual blur, for example in the field of high dynamic range video reconstruction.

Chapter 6

Eye-Tracking Head-mounted Display

Contents

Being able to detect and to adapt to gaze direction facilitates many new ways to enhance digital displays. Especially for head-mounted displays, knowing in real-time where the user is looking enables providing for a much improved viewing experience, e.g. by realistically animating virtual avatars or enhancing depth perception by gaze-aware depth-of-field rendering. In addition, gaze-contingent video coding and foveated rendering strategies enable bandwidth reduction, higher frame rates and overall better rendering performance. Head-mounted displays especially pose high demands on spatial resolution and constantly high refresh rates and therefore significantly benefit from these techniques.

In this chapter a novel modular HMD design with integrated binocular eye tracking is presented. The VR headset enables directly adapting the displayed frames to the user's current gaze direction.

6.1 Introduction

Virtual Reality (VR) has become a well-established field in research and industrial applications, e.g., for simulations, scientific visualization, or gaming. Previously, high hardware costs prevented a wide-spread application and development. But recent advances in the mobile device market lead to high-quality, low-cost virtual reality hardware (*Oculus Rift, HTC Vive, Sony Playstation VR*, etc.). These low-weight, low-latency head-mounted displays (HMDs), in combination with a wide field of view (FOV), enable experiencing *immersion* and *presence* within a virtual environment like never before. Future developments of HMDs will include even higher resolution displays, higher refresh rates, and wider FOVs [Abr14].

Commodity HMDs mostly use fixed hardware setups. Unfortunately, preconfigured HMDs are often difficult to parameterize for individuals, e.g., to account for differing interocular distances both in horizontal and vertical direction (previously often ignored and known as Hypertropia [DFRR10]). Further, existing software calibration is often unsatisfactory and cumbersome with current HMDs. This limitation can lead to non-frontal relative positioning of the eyes and non-converging lenses inside the HMD, resulting in reduced perceived sharpness, and an increased likelihood of motion sickness and headaches for the user. The wide adoption of VR equipment justify to investigate methods to simplify calibration and to improve the experience for every user. Here, analyzing user behavior in virtual environments can deliver many insights: What is drawing attention? What emotional response results from certain content?

For a desktop setup similar questions are usually investigated involving an eye tracker (measuring pupil size for emotions or focus points of interests on the screen). Unfortunately, when using an HMD setup, the integration of eye tracking is not straightforward, and existing solutions are not convenient for commodity HMDs.

Stationary solutions for eye tracking are state-of-the-art with regard to tracking quality and are mostly applied to estimate scan paths (fixations and saccades) [HNA+11, Sch14]. The user's head is locked in position using a rigid positioner while cameras record the eyes. While the systems are very accurate and provide high tracking sample rates, the fixated viewing position is not an option

for immersive VR where even small head movements lead to drift if recalibration is not frequently performed.

Mobile eye-tracking solutions overcome the fixation restriction. In this case, an integration into a headgear or special-glasses frame enables free head movement (e.g., SMI Eye Tracking Glasses, Arrington Research 3DViewPoint™, Biopac Systems, Inc. HMD). However, due to its smaller form factor it is significantly more ambitious to integrate such a solution into an HMD: The tracking relies on a camera whose position is constrained by the HMD lenses and lens holders, which can partially block the view. Hence, a point for the camera right below the eyes is chosen where precision is, unfortunately, non-uniform. An alternate, more-frontal placement inadvertently reduces the FOV, which is often not an option because the feeling of immersion only starts at a horizontal FOV of 80°and increases quickly until 110° [Abr14, BKLJP04]. Such an eye tracking procedure is further complicated since typical corneal-reflection-based eye-tracking algorithms [HJ10] are not applicable as they would produce disturbing reflections on the lenses.

This chapter addresses these limitations and works towards gaining more insight into the VR experience. An affordable, drift-free and binocular eye-tracking solution is presented which is usable within the limited space of current HMD hardware designs without FOV reduction (Fig. 6.1). Throughout this chapter, it is shown how to overcome the challenges involved in designing such a VR system and solve several other issues, for instance calibration and adaptation to the user.

Specifically, the contributions of this work are:

- a personalizable lens positioning system (horizontal and vertical) for HMDs and an integration of an unobtrusive camera setup for eye tracking in a lens-based HMD based on infra-red lighting (Sec. 6.2);
- a model-based gaze estimation algorithm and calibration procedure to adjust the system to the user (Sec. 6.3);
- a robust monocular pupil-tracking algorithm which can deal with partial eye occlusions by lens holders and eye lid (Sec. 6.4);

To show the potential of (binocular) eye tracking in HMDs a variety of novel applications are presented, such as gaze-contingent level-of-detail, accommodation simulation, gaze map creation, and realistic gaze control of virtual characters. In general, these applications illustrate the ability of the system to perform psychophysical experiments and to extend the experience in immersive environments (Sec. 6.5). The proposed system is validated by an objective comparison with a state-of-the-art pupil-tracking algorithm for near-field eye-trackers [LWP05] as well as by a user evaluation (Sec. 6.6). The limitations of the current setup and future work are discussed (Sec. 6.7) before concluding the chapter (Sec. 6.8).

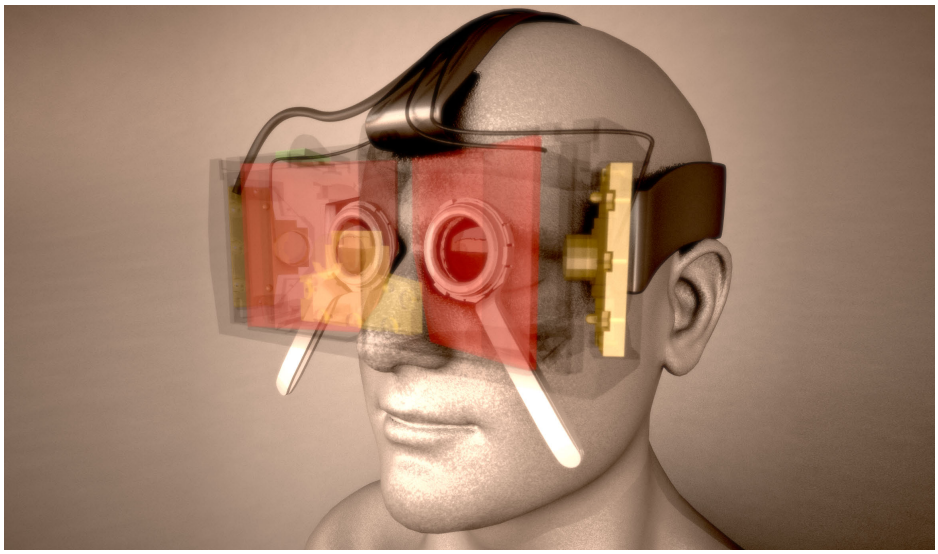

Fig. 6.1 Prototype visualization. A rendering of the proposed self-contained eye-tracking head-mounted display. Based on a system of dichroic mirrors (red), lens units illuminating the eye balls in the infrared spectrum (white) and tracking cameras (yellow) the device captures the user's eye motion for binocular eye-tracking while he is fully immersed in the virtual world.

Related VR Headsets The success of the Rift™ HMD from Oculus VR led to a renewed interest in VR for the consumer market. The most-evolved HMDs in this low-cost sector, Rift (Oculus VR) and HTC Vive, offer a display resolution of at least Full-HD as well as positional and rotational tracking. Eye tracking is a natural next step and gained much attention in the research and development sector (e.g., FOVE Inc., Arrington Research, ASL Eye-Trac 6, SR Research, or Senso Motoric Instruments (SMI)). Even though first attempts have been undertaken in the year 2000 [DSR+00], current prototypes are still far from being consumer-ready[1]. One major cost factor are the miniature cameras and specialized digital processors for tracking at high speed. While the interior design of these Eye-tracking HMDs (ETHMD) is mostly kept secret, the comparatively low vertical FOV suggests that the camera is placed inside the user's FOV, occluding part of the display. In contrast, the proposed eye-tracking HMD setup has here several benefits. It is a low-cost solution (approximately 450$ in hardware), and preserves the full FOV of current state-of-the-art HMDs. Closest to the described design is the EyeSeeCam [SVV+09]. This wearable eye tracker is used to align the focus of an external camera and the user in real-time for medical applications, surgery, or behavioral sciences. Similarly, the ETHMD described in this chapter uses dichroic mirrors to reflect infrared light from the eyes back to the cameras located outside the FOV. The custom-built EyeSeeCam relies on traditional eye-tracking algorithms, and is much more expensive. For the presented ETHMD, additional challenges had to be solved such as partial occlusion by the lens holders and view distortion by the lenses.

[1] 15,000$ for SMI's eye tracker for the Oculus Rift, 11/2014, http://newatlas.com [Upg14]

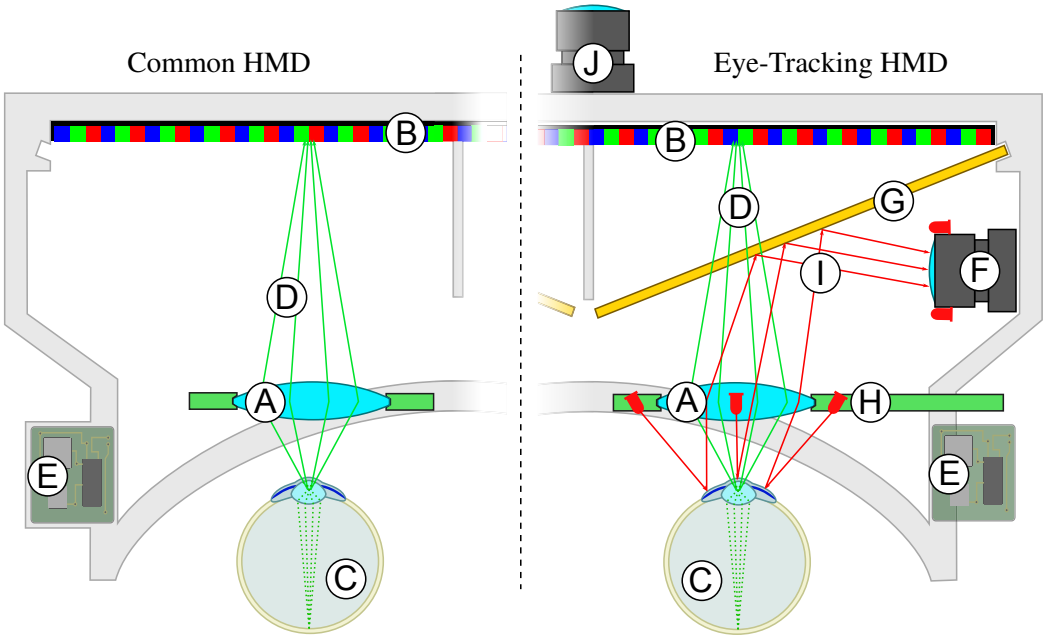

Fig. 6.2 HMD design comparison. Common HMDs setup (left): converging lens (A), display (B), eye ball (C), visible light (D), head orientation tracker (E); The proposed system adds (right): eye tracking camera (F), dichroic mirror (G), lens holder with infrared LEDs (H), infrared light (I), positional tracking camera (J).

6.2 Eye-Tracking HMD

The following sections present the low-cost, low-weight and personalizable design, and describe details of each of the HMD's main components for immersive VR with unobtrusive eye tracking.

6.2.1 Device Construction

General The important elements of the ETHMD are visualized in Fig. 6.2. The working prototype is depicted in Fig. 6.3a–b. The basic setup resembles a classic HMD with a converging lens per eye (A) to focus the view on the display (B). The difference lies in a pair of infrared cameras to the side of the body case (F), dichroic mirror (G) and a circular LED-light array around the adjustable lens holder (H) to illuminate the eye (C). Reflected infrared light passes through the converging lens and is reflected towards the camera via the tilted dichroic mirror (G). Light from the display (D) passes unhindered towards the eye. An additional front camera (J) outside on the HMD is used for markerless positional tracking. An integrated inertial measurement unit (E) is used in addition for head orientation [DWB06]. The electronic components are wired to a single harness connected to an external box with the display controller, a micro computer for orientation tracking, and the LED power supply [ARD15].

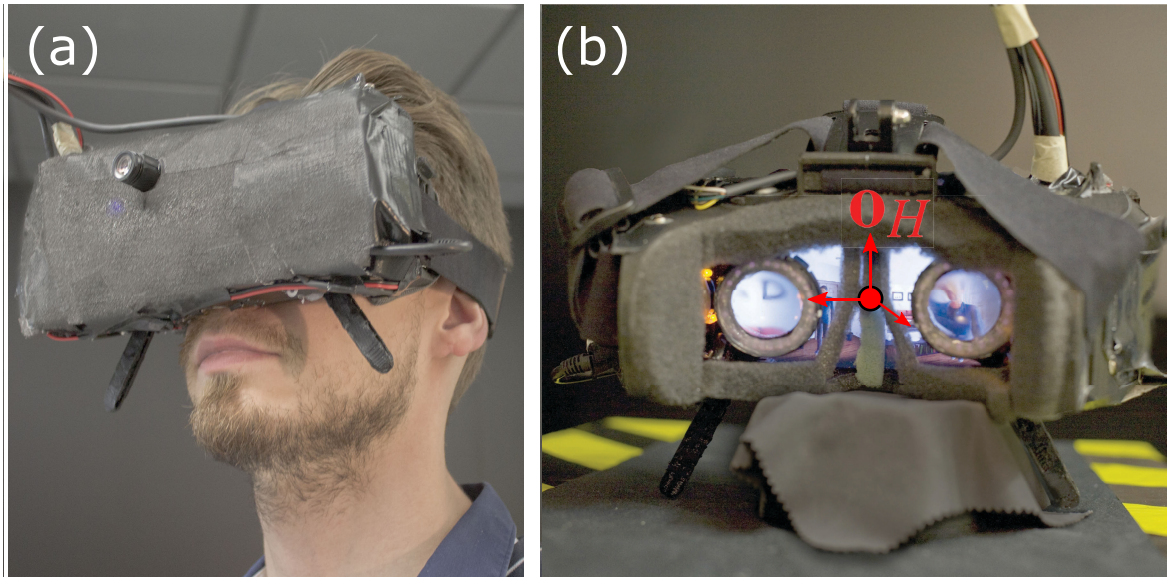

Fig. 6.3 HMD design and assembly. User wearing ETHMD prototype (a), and body frame (b).

Body Case The body case encapsulates all internal components (Fig. 6.2). A central barrier with a gap for the nose divides the display (C) into two disjoint symmetric parts, one for each eye. The case closes firmly and tight around the eyes. Foamed material avoids exterior stray light. It is covered with comfortable tissue, except at the nose tip to enable normal breathing. The dimensions of the body are adjusted to the average head size of human adults [NSMB+12].

Display The integrated 5.6" LCD display (2560×1440 pixel resolution) works at a refresh rate of 60 Hz. As indicated before, display controller and display are separated which reduces HMD weight.

Converging Lenses The use of converging lenses, salvaged from an Oculus Rift (DK1), enables increasing the perceived field of view and adapts the focal distance to a comfortable distance [OCU15]. Compatibility to the Oculus Rift is maintained. The prototype offers a horizontal field of view of 86°per eye. Dedicated controllers are provided to adjust the position of the lenses in both horizontal and vertical direction for optimal lens placement (Fig. 6.1). Compared to a traditional HMD with interchangeable lens cups, the proposed design makes possible more flexible and precise adjustments for varying head anatomy. For calibration, a circular IR-reflecting ring is located on the backside of the lens holders.

Dichroic Mirrors Two dichroic, planar mirrors (also known as *hot mirrors*) are used which reflect light at wavelengths longer than 730 nm (infrared), while short wavelengths (< 720 nm) are entirely transmitted. They redirect infrared light reflected by the IR-illuminated eyes towards the integrated cameras, which allows tracking the gaze without obscuring the field of view of the user.

The dichroic mirrors have a size of $80 \times 80 \times 2$ mm with central cutouts for the nose and an inclination angle of 19.5° along the vertical axis. The angle is a tradeoff between space and optimal view on the eye (45° inclination). Higher inclination angles would increase the necessary screen distance and, thus, screen size and weight. A smaller inclination angle, on the other hand, leads to a partly occluded view at the eye which need to be dealt with during pupil tracking.

Infrared Illumination Unit Twenty-five infrared LEDs mounted on a ring circuit, uniformly illuminate each eye from all directions (Fig. 6.4). The ring has an inner diameter of 37 mm and width of 1.5 mm to minimize the overall lens holder size. The LEDs radiate at a wavelength of 830 nm and a wide angle of 140°. The infrared light enhances the contrast between pupil and iris, but is outside the visible spectrum, thus, invisible to the user. Safety of the user is ensured with regard to the exposure of infrared radiation (Sec. 6.2.2).

Eye Tracking Cameras For binocular eye tracking, two low-cost cameras with a fixed diagonal field of view of 56°are used to record the user's eyes. The infrared blocking filters of the cameras are removed. Instead, a long-pass filter, consisting of a single layer of a raw film negative and blocking all but infrared light, is inserted. The cameras are mounted fixed in the HMD (Fig. 6.2F), and record at 640×480-pixel resolution in grayscale at 75 Hz. The cameras feature a delay of 13 ms due to the internal image processor. The sampling rate suffices to track fixations and smooth pursuit eye movements.

Head Tracking For viewpoint estimation in a virtual environment, the rotational and translational location of the HMD are required. An orientation sensor integrated into the HMD and a head-mounted front camera performs positional tracking. This combined setup is inexpensive and provides 6-degrees-of-freedom head tracking with sufficiently high precision and low latency.

Orientation (Yaw-Pitch-Roll) is tracked by an inertial measurement unit (IMU) holding multiple sensors connected to an Arduino microcontroller board. The IMU consists of an accelerometer, a gyro sensor and a digital motion processor (DMP). The update rate is set to 200 Hz to avoid any noticeable delay when moving the head in order to reduce motion sickness.

The DMP supports automatic self-calibration, and the angular drift of the IMU is less than 1° per minute, which is sufficient for longer usage. The positional tracking of the IMU can suffer from an integration error over time, resulting in accumulated drift. Over short periods of time, however, the IMU delivers sufficiently precise data. The used tracking solution combines the low-latency IMU output with markerless camera tracking, which results in robust, low-latency positional tracking with good precision. The pose estimation of the HMD front camera in world space is based on SLAM-feature tracking, implemented in the Metaio SDK [DWB06, MET14]. Pose estimation proceeds as follows: The world frame is oriented and positioned automatically after a few seconds of feature initialization. Features are then detected and refined adaptively over time during tracking. Since world scale cannot be estimated from the tracker, it is automatically set during the initialization phase

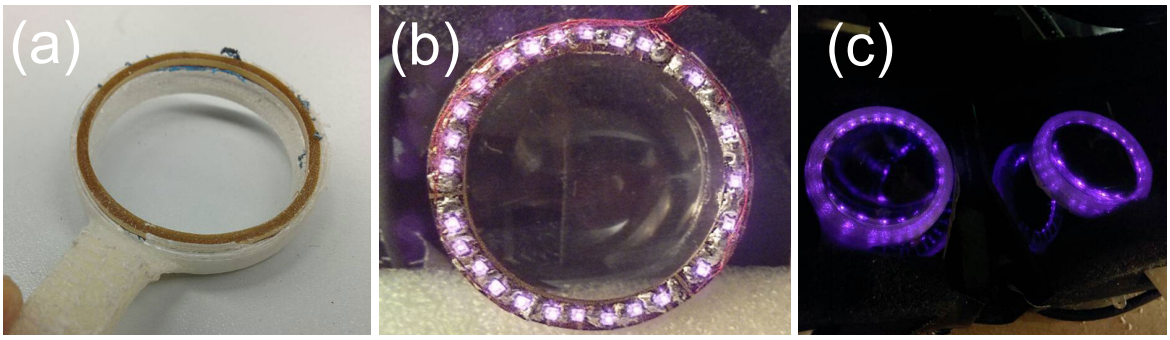

Fig. 6.4 Eye-illuminating lens holder. 3d-printed lens holder with manufactured circuit board (a), working infrared SMD-LED array (b), illumination units inside the HMD (c).

such that the camera-tracker results are consistent with the velocity measured by the IMU. Positional tracking takes $\approx$ 23 ms (13 ms for frame transmission, 10 ms for pose estimation).

6.2.2 Safety Analysis

Strong infrared light can potentially damage the retina, but the level of infrared radiation of the proposed ETHMD setup is not harmful. Following Mulvey et al. [MVS+08], for a single LED the relevant solid angle is given by

$$\Omega = \frac{\pi \cdot r^2 \cdot \cos(\alpha)}{d^2}, \tag{6.1}$$

where α is the angle between the optical axis and the vector from the LED to the center of the cornea and r the pupil radius (fully dark-adapted approximately 0.4cm). Ω in Eq. 6.1 is largest at a distance between pupil and LED of $d = 2.1$ cm resulting in a solid angle of $\Omega_{max} = 0.0499sr$.

The power of each LED is given by $P_{\text{LED}} = 2.5\ mW/sr$. Hence, the total maximal irradiance for our radially-symmetric $n = 25$ LED setup per eye is

$$E_{\text{max}} = P_{\text{LED}} \cdot n \cdot \Omega_{max} \approx 3.116\ mW/cm^2 \quad , \tag{6.2}$$

which is well below the recommended daily maximum irradiance of 10 mW/cm^2 for a wavelength of 700 - 3000 nm per eye [MVS+08].

6.3 Calibration

This section describes calibration procedures for the different HMD components (Sec. 6.3.1) and the user-specific calibration (Sec. 6.3.2). Both are required for precise eye tracking, gaze estimation, and personalized adjustments. It is an important step in fitting the device to the user which, ultimately, leads to a better VR experience. The setup is described for one eye. The second eye is handled equivalently. The eye-tracking implementation is described in Sec.6.4.

6.3.1 HMD Calibration

To avoid motion sickness and to create a convincing 3D impression, precise knowledge about each component in our HMD projection chain is required, i.e. the relative position and orientation of the eye-tracking camera, the dichroic mirror, the lens and lens holder, as well as the refractive properties of the converging lens and the intrinsic parameters of the eye-tracking camera. As a reference point $\mathbf{o}_H$ for all components of the HMD, the horizontal center of the HMD's front-most point is used, Fig. 6.3 right.

Eye-Tracking Camera Calibration For calibration both intrinsic and extrinsic parameters of the eye-tracking camera are estimated. Intrinsics are derived via the technique by Bouguet [Bou10]. Providing image resolution and sensor size is sufficient to transform a recording of a checkerboard or circular pattern of known size into focal length, principal point, as well as radial and tangential lens distortion.

The extrinsic camera parameters are derived during component assembly as follows. Before the dichroic mirror is inserted into the body case the display is covered with a checkerboard calibration pattern which is carefully adjusted to align with the edges of the body case. The eye-tracking camera records the pattern, and the extrinsic parameters are derived in relation to the pattern. The same CAD model which is used to print the body case is used to relate the extrinsic camera parameters to $\mathbf{o}_H$, the coordinate system of the HMD [HZ03]. For validation, the captured image is compared with a rendered version of the checkerboard using the derived camera parameters. The reprojection error is less than 3 pixels and can certainly be further reduced in an industrial production setting.

Mirror Calibration After the camera has been calibrated, the dichroic mirror is inserted and calibrated. The mirror is covered with a carefully aligned calibration pattern to match it later to the CAD model, and to capture it from the eye-tracking cameras. Performing the same calibration procedure as for the cameras yields the camera parameters in relation to the mirror position, and vice versa. This relative mirror position is then transformed to $\mathbf{o}_H$. Again, the correctness of the derived parameters is validated by rendering the checkerboard and comparing it with the captured image. In the proposed ETHMD prototype, the rotation angles of the mirrors are found to be $\sim 18.9^\circ$ and $\sim 19.5^\circ$ for the left and right mirror, respectively. The slight asymmetry was due to a fabrication imperfection when printing the HMD.

Lens Reconstruction An accurate geometric model of the aspheric lens as well as the index of refraction (IOR) are required to support user calibration later on. For the used lenses, details about optical properties have not been available and had to be reconstructed. As this is the situation for most HMD lenses, the performed lens reconstruction is described in the following.

Lens Geometry To avoid complicated reconstruction of a transparent surface, the lens surface is artificially colorized with ink which creates a set of discriminative features. The input images are

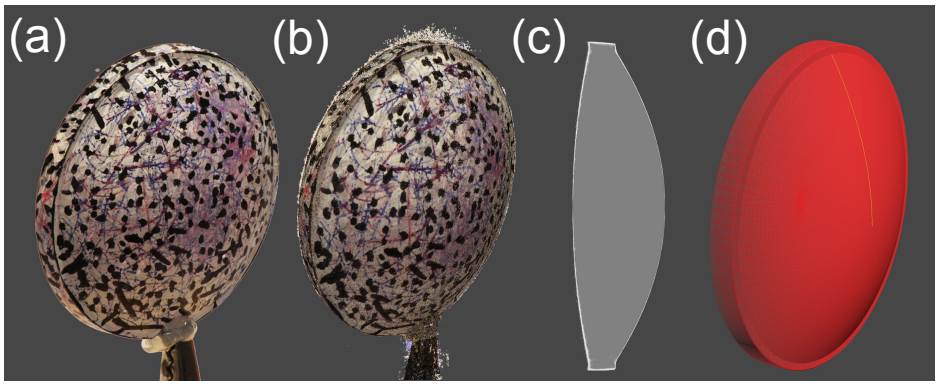

Fig. 6.5 Lens reconstruction. Converging lens with artificial surface features (a), reconstructed 3d point cloud (b), derived lens profile (c), reconstructed lens (d).

captured at high quality and resolution using a DSLR camera (Fig. 6.5a). Then, a lens-surface point cloud based on different input views is reconstructed (Fig. 6.5b) [AGI14]. As a point cloud may contain holes, a parametric lens model is derived (Fig. 6.5c) as follows. The approach reasonably assumes a disc-like and radially symmetric shape. The mean positional vector μ of the point cloud and the eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ provide a convenient coordinate space for the lens reconstruction as μ is equal to the center of the lens and together with $\mathbf{e}_3$ describes the rotation axis $\mathbf{r} = \mu + t\mathbf{e}_3$. Because of the symmetry assumption, only the 2D profile needs to be derived (Fig. 6.5c). It can be conveniently described by two second-order polynomials for the front and back curvature. Each point of the point cloud around $\mathbf{r}$ is rotated onto the plane centered at μ and spanned by $\mathbf{e}_1, \mathbf{e}_3$. Then two second-order polynomials are fitted to the point cloud data, one for the front-facing points and one for the backfacing points [Moi11]. This approach also increases robustness as the symmetry assumption leads to better use of point cloud redundancy. The lens is then modeled from this parameterized profile (Fig. 6.5d).

Index of Refraction Since the index of refraction (IOR) of the lens is wavelength-dependent, it is estimated for infrared light and for the visible spectrum (Fig. 6.2). The following procedure is the same for both, only the recording camera is exchanged. An analysis-by-synthesis approach is applied based on the lens's geometric properties. First, a front view of a circular calibration pattern is captured (Fig. 6.6a). It has an outer diameter of 50mm at a known distance and the cameras are calibrated as before. After adding the lens between camera and pattern, several images are captured at different, known distances between pattern and lens (Fig. 6.6b). Then, the IOR is optimized by comparing to the recorded camera images synthetically rendered scenes of the lens and calibration pattern (Fig. 6.6c) using the physically-correct and wavelength-dependent Maxwell Renderer [MAX15] . Two IOR values are derived: $N_I = 1.472$ for $\lambda = 950nm$ and $N_V = 1.515$ for $\lambda = 560nm$ which are typical values for materials like Acrylite, Lucite or Plexiglass.

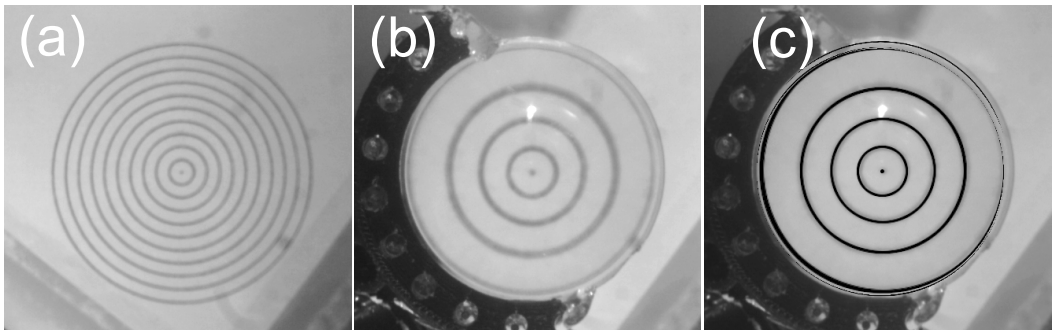

Fig. 6.6 Refractive index estimation. Calibration object (a), ground truth refraction through the lens (b), calibration pattern rerendered on top of ground truth (c).

6.3.2 User Calibration

Most components of the ETHMD can be calibrated during assembly (Sec. 6.3.1). User-specific components, such as the lens holder position, interpupillary distance, and eye-to-lens distance need to be estimated for every user separately. This is important to provide a natural 3D impression and meaningful eye-tracking results, because these values are essential to predict the virtual viewpoint which can otherwise only be roughly estimated. The components of the gaze simulation model are visualized in Fig. 6.7a. First, the user adjusts the lenses parallel to the screen for a frontal view when looking straight ahead. Next, the lens distance is adjusted until the screen appears sharp.

Lens-Holder Localization To detect the lens-holder position, and, hereby, the lens's position, the white IR reflecting ring on the backside of the lens holder are used (Fig. 6.7a). Additional infrared LEDs are located around the eye-tracking camera solely for illuminating the ring (Fig. 6.2 (F), red LEDs at the camera). When turning off the screen and the interior LED ring, the lens holder can be detected by thresholding the image captured by the eye-tracking camera. Then, its center and eccentricity is derived [FF+96]. The 3D position and orientation of this ring is estimated again via an analysis-by-synthesis procedure; a model of the ring is rendered and its position and rotation are iteratively optimized via a gradient-descent approach based on the difference between the ellipse centers, size and eccentricity which proves fast and accurate.

Eye Calibration Next, the eye's distance to the lens is estimated. The main problem is that a view of the eye does not provide useful information regarding scale as eye sizes differ. Further, the view might be distorted in complex ways by the converging lenses. Analysis-by-synthesis again helps. The LED ring in the lens holders produces characteristic reflections on an eye, also known as glints (Fig. 6.7 b–c). The reflections can be used to determine the distance between lens and eye. Nonetheless, to make this step possible a physically plausible eye model is required.

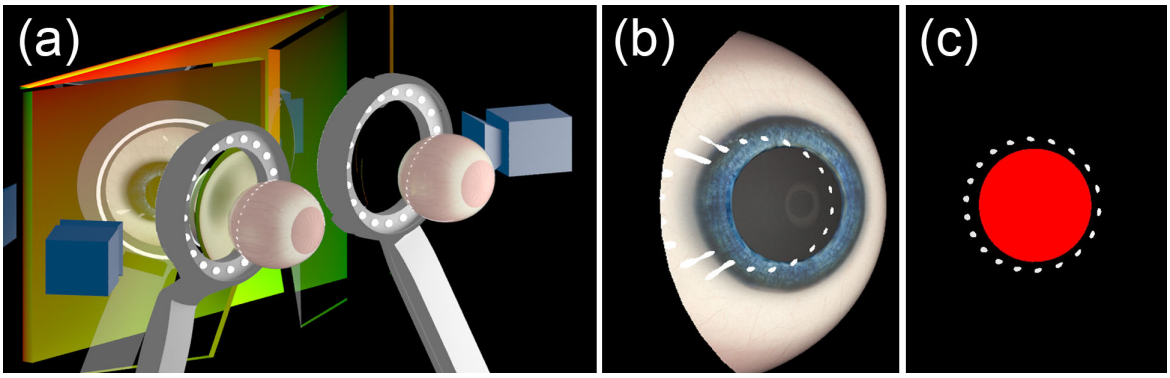

Fig. 6.7 Simulated gaze model. Virtual gaze setup (a), synthetic eye (b), glints and pupil mask for characteristic gaze (c).

Eye Model The eye ball of a healthy adult human has a quite consistent shape. The main part can be modeled as a sphere rotating around its center with a diameter of 24mm and only few individual deviations (Gaussian distribution with a standard deviation of $\pm$1mm) [AKLA11, p. 75]. The cornea forms an additional spherical surface above the sclera with a smaller radius of 7.8mm (Chapter 2.2). The direct light reflected off the sclera produces the most prominent glints. The IOR of the cornea is set to $N_D = 1.2$ and the eye fluids to $N_D = 1.276$ [AKLA11].

Eye Registration When a user looks along the optical axis of the eye-tracking camera (taking the reflection from the dichroic mirror into account) the glints form an almost perfect circle on the sclera. For any other view direction this circle is distorted. The shape of the ring of glints can, thus, be used in a feedback loop to guide the user's view towards the optical axis of the eye-tracking camera (Fig. 6.7b–c). Towards this goal, the user is asked to focus on a marker on the screen. From the camera image the pupil-center position c_p is estimated by using the pupil-tracking algorithm described in Sec. 6.4. In addition, the glints are extracted from the eye-tracking camera frame using a simple brightness threshold $t_G = 0.9$. The resulting "glint mask" is used to fit an ellipse [FF+96]. The glint-ellipse center in pixel coordinates is declared as c_g.

The marker is then moved and the ellipse is evaluated again from the next camera frame. The movement of the marker is given by the difference of glint-ellipse center and pupil-center position, $\alpha(c_g - c_p)$. The process is computationally cheap and α can be small, which lets the marker smoothly move over the screen until the algorithm converges.

Then, the eye-lens distance and the absolute 3D position of the eye are derived from the eye model. In practice, the characteristic positions where the glints are as circular as possible are rendered for each eye distances, eye positions and each possible lens holder configuration. The transition distance from one (virtual) eye configuration to the next configuration is set to 0.5mm. For each rendered glint image the diameter and center of a fitted ellipse is recorded. The result is a Look-Up table which

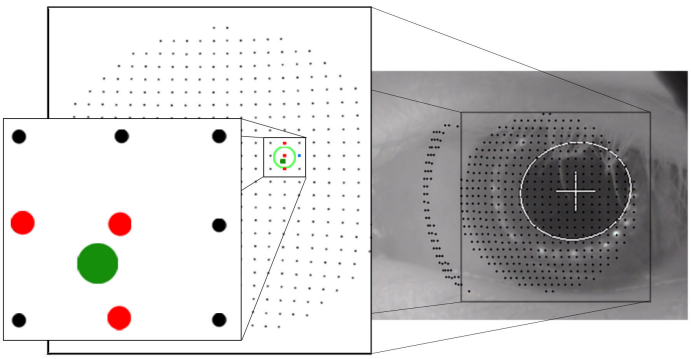

Fig. 6.8 Gaze mapping. Barycentric interpolation for pupil-to-screen mapping (green: current pupil position, black: precomputed positions, red: closest samples used for interpolation).

allows calibrating for the eye position quickly by linear interpolating between the four closest eye positions for the previously estimated lens holder configuration.

Gaze Calibration Finally, for gaze estimation, a mapping from *pupil center position* in the eye-tracking camera to *screen position* is needed. This mapping is precomputed using the virtual HMD model which is configured with the derived calibration values and estimated eye position. A ray from one pixel representing a detected pupil center is cast via the dichroic mirrors through the lens towards the eye. By construction, this ray has to intersect with the eye at the pupil center. Thus, eye-tracking camera pixels can be mapped to eye rotation angle. This mapping is precomputed for approximately 1300 virtual eye rotations per eye covering the full motion range of the human eye (Fig. 6.8, black and red dots). Similarly, the eye rotation can be used to determine screen position by computing the light path from the eye through the lens onto the screen (Fig. 6.2 green light paths). Using barycentric interpolation, each detected pupil-center position c_p in the eye-tracking camera (Fig. 6.8, green dot) can be mapped to a view vector $\vec{v}$ and to a pixel position on the display c_s.

6.4 Pupil Tracking

The implemented gaze estimation algorithm (Sec. 6.3) relies on detecting the current pupil position in the eye-camera image. The pupil extraction is described in this section for which noise, (partial) occlusion by the eye lid or lashes and dust or smears on the lens or mirror need to be handled. Since the off-axis illumination results in a *dark pupil*, pupil regions result in low pixel intensities, and differ significantly from the high amount of reflected infrared light from the sclera and iris. For pupil tracking (Alg. 1), grayscale camera images $\mathbf{I}$, normalized to $[0,1]$ are used as input.

A binary mask $\mathbf{M}_L$ indicating pixels belonging to the lens is obtained during the calibration step and is applied to each input frame. First, for real-time performance and robustness, it is determined whether the eye is closed, open or halfway closed. Each configuration is dealt with separately (Alg. 1).

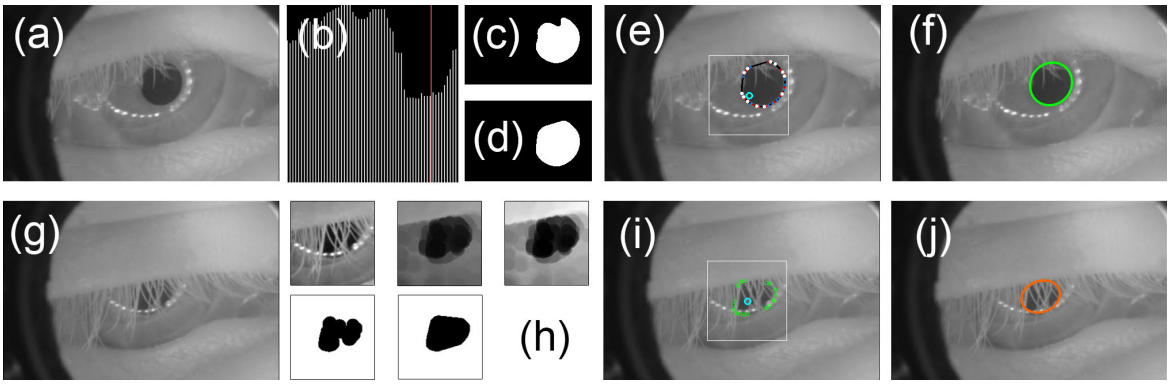

Fig. 6.9 Pupil detection pipeline. Top row: Visible pupil case, Alg. 4. Captured image (a), histogram for threshold estimation (b), pupil binarization (c), pupil closing (d), contour filtering (e), pupil ellipse fitting (f). Bottom row: Partially occluded pupil case, Alg. 5. Captured image (g), pupil filtering and binarization (h), contour point filtering (i), pupil ellipse fitting (j).

Algorithm 1 Pupil Tracking $(\mathbf{I}, \mathbf{M}_L)$

$\widetilde{p} \leftarrow$ Approximate Pupil Position $(\mathbf{I}, \mathbf{M}_L)$ ▷ Alg. 2
if $\mathbf{I}(\widetilde{p}) > t_{visibility}$ **then**
    Eye is closed
else
    $\theta \leftarrow$ Compute Pupil Occlusion $(\mathbf{I}, \mathbf{M}_L, \widetilde{p})$ ▷ Alg. 3
    if $\theta < t_{occlude}$ **then**
        $\{p, e_x, e_y, \phi\} \leftarrow$ Detect Visible Pupil $(\mathbf{I}, \mathbf{M}_L)$ ▷ Alg. 4
    else
        $\{p, e_x, e_y, \phi\} \leftarrow$ Detect Occluded Pupil $(\mathbf{I}, \mathbf{M}_L)$
        ▷ Alg. 5
    end if
end if
return $\{p, e_x, e_y, \phi\}$

Algorithm 2 Approximate Pupil Position $(\mathbf{I}, \mathbf{M}_L)$

$p_{cum} \leftarrow (0,0) \quad w_{cum} \leftarrow 0$
for $p \in \mathbf{M}_L$ **do**
    $w \leftarrow (1 - \mathbf{I}(p))^{\gamma}$
    $p_{cum} \leftarrow p_{cum} + p \cdot w$
    $w_{cum} \leftarrow w_{cum} + w$
end for
return $\widetilde{p} \leftarrow p_{cum} / w_{cum}$

Algorithm 3 Compute Pupil Occlusion $(\mathbf{I}, \mathbf{M}_L, \widetilde{p})$

$\mathbf{M}_{Glints} \leftarrow \left\{ p \in \mathbf{M}_L \mid \mathbf{I}(p) > t_G \right\}$ ▷ Glint Mask
$\mathbf{M}_{Glints} \leftarrow F_{\text{Dilate}}(\mathbf{M}_{Glints})$
$\mathbf{I}_{NoGlints} \leftarrow F_{\text{Inpainting}}(\mathbf{I}, \mathbf{M}_{Glints})$ ▷ Glints removed
$\mathbf{I}_{MinMax} \leftarrow F_{\text{Min}}(\mathbf{I}_{NoGlints}, k_{MinMax})$ ▷ Eye lashes removed
$\mathbf{I}_{MinMax} \leftarrow F_{\text{Max}}(\mathbf{I}_{EyeLashes}, k_{MinMax})$
$\mathbf{I}_\Delta \leftarrow |\mathbf{I}_{MinMax} - \mathbf{I}_{NoGlints}|$
$\mathbf{M}_{ROI} \leftarrow F_{\text{circMask}}(\mathbf{I}, \widetilde{p}, r_1) \cap \mathbf{M}_L$
$\mathbf{I}_\Delta \leftarrow \text{Normalize}(\mathbf{I}_\Delta \cap \mathbf{M}_{ROI})$
$m_1 \leftarrow \sum_{p \in \mathbf{M}_{ROI}} \mathbf{I}_\Delta(p) / |\mathbf{M}_{ROI}|$ ▷ First Metric
$\mathbf{I}_{EyeLashes} \leftarrow F_{\text{(Tonal Correction)}}(\mathbf{I}_{NoGlints}, [0.4, 0.7])$
$\mathbf{M}_{ROI} \leftarrow F_{\text{circMask}}(\mathbf{I}, \widetilde{p}, r_2) \cap \mathbf{M}_L$
$m_2 \leftarrow \sum_{p \in \mathbf{M}_{ROI}} \mathbf{I}_{EyeLashes}(p) / |\mathbf{M}_{ROI}|$ ▷ Second Metric
$\theta \leftarrow (m_1 \cdot w_1 + m_2 \cdot w_2) \cdot 0.5$ ▷ Combined Metric
return θ

Approximate Pupil Position To make a fast guess of whether the eye is closed or not the algorithm approximately locates the pupil position as depicted in Alg. 2. Within the lens mask $\mathbf{M}_L$ a weighted average of all pixel positions p is accumulated. Each pixel p contributes with a weight w determined by $(1-\mathbf{I}(p))^\gamma$ with $\gamma = 10$. Hence, darker pixels (higher likelihood to be the pupil) will contribute more. The weighted-average position is the initial pupil-position guess $\widetilde{p}$.

Occlusion Estimation If the intensity in a 70×70 pixel wide window around the initial pupil position is above the threshold $t_{visibility} = 0.4$, the eye is regarded as being closed. If the eye is not completely closed, the tracking strategy is further refined by classifying the eye as either completely visible or partially occluded. The amount of occlusion is defined by two measures m_1 and m_2 (Alg. 3). While not being sufficient on their own, the combination is significantly more robust. The first, m_1, estimates the presence of eye lashes. The second, m_2, tries to detect the amount of non-pupil pixels. Based on their combination, the eye is declared visible or partially visible, and the corresponding detection algorithms are run.

To compute m_1 and m_2, glints are removed by inpainting all pixels with an intensity higher than $t_G = 0.9$ [BBS01]. These appear especially in the 1.5 – 2 mm transition zone of the curvature of the sclera and the curvature of the corneal surface that forms an external and internal surface groove (*scleral sulcus*) [AKLA11, p. 75]. For more conservative results the inpainted area $\mathbf{M}_{Glints}$ is slightly dilated.

Then, eye lashes occluding the pupil are detected in the resulting image $\mathbf{I}_{NoGlints}$. Computations are restricted to a small area $\mathbf{M}_{ROI}$ of radius $r = 35$ pixels around $\widetilde{p}$. Then a morphological opening filter is applied (minimum before maximum filter) to $\mathbf{I}_{NoGlints}$ with a kernel size $k_{MinMax} = 13$, removing

Algorithm 4 Detect visible pupil $(\mathbf{I}, \mathbf{M}_L)$

$\mathbf{I}_{BP} \leftarrow 1 - \mathbf{I}$ ▷ Inverts to bright pupil image
$\mathbf{H} \leftarrow \text{Histogram}(\mathbf{I}_{BP}, \mathbf{M}_L)$
$\mathbf{H} \leftarrow F_{\text{Median}}(\mathbf{H}, k_{HistMedian})$
$h \leftarrow \text{findGrayvalueOfBrightestLocalMinimum}(\mathbf{H})$
$\mathbf{M}_{PupilSeg} \leftarrow \left\{ p \in \Omega \mid \mathbf{I}_{BP}(p) > h \right\}$
$\mathbf{B} \leftarrow \text{Blob Detection}(\mathbf{M}_{PupilSeg})$
$\mathbf{b} \leftarrow \text{argmax}_{\tilde{b} \in \mathbf{B}} \text{HullArea}(\tilde{b})$
if $\text{HullArea}(\mathbf{b}) < t_b \cdot \sum_{\tilde{b} \in \mathbf{B}} \text{HullArea}(\tilde{b})$ **then** ▷ Merge blobs
    $\mathbf{b} \leftarrow \mathbf{b} \cup \left\{ \tilde{b} \in \mathbf{B} \mid \|\text{Centroid}(\mathbf{b}) - \text{Centroid}(\tilde{b})\| < d \right\}$
end if
$\mathbf{C} \leftarrow \text{Convex Hull Contour}(\mathbf{b})$
$\mathbf{C} \leftarrow \text{Remove Close Points}(\mathbf{C})$
$\mathbf{C} \leftarrow \text{Remove Colinear Points}(\mathbf{C})$
$\left\{ p, e_x, e_y, \phi \right\} \leftarrow \text{Ellipse Fit}(\mathbf{C})$
return $\left\{ p, e_x, e_y, \phi \right\}$

finer structures, such as eye lashes. The first term m_1 is then defined as the sum of absolute intensity values of the difference image $\mathbf{I}_\Delta = |\mathbf{I}_{MinMax} - \mathbf{I}_{NoGlints}|$.

The second term aims at estimating the number of non-pupil pixels, which are brighter. To this extent, the gray value range $[0.4, 0.7]$ in $\mathbf{I}_{NoGlints}$ is linearly mapped to the range $[0, 1]$. Other values are clamped accordingly. The second term m_2 is then defined as the sum of the resulting intensities inside $\mathbf{M}_{ROI}$. Both terms are combined into the final occlusion score $\theta = 0.5 \cdot (1/3 m_1 + 2/3 m_2)$. If $\theta < t_{occlude} = 0.3$ the eye is considered visible, otherwise partially occluded. The corresponding detection algorithm is applied.

Visible Pupil In the following, localization of a visible or moderately-occluded pupil is described (Alg. 4). The algorithm builds upon the observation that pupil pixels in comparison to their surrounding are well separated in an image histogram (Fig. 6.9a–b). Thus, a histogram $\mathbf{H}$ is computed on the inverted input image $\mathbf{I}_{BP} = 1 - \mathbf{I}$ with 64 bins. A median filter of size $k_{HistMedian} = 2$ removes outliers. Marking pixels brighter than a threshold h separates the pupil well. Following the observations, h is set to be the grayvalue belonging to the brightest local minimum within $\mathbf{H}$ (Fig. 6.9b, red bar in histogram).

Next, the goal is to clean up the derived pupil pixels and to perform blob detection ($\mathbf{B}$) to find connected components. Inspired by Chen et al. [CE14], the convex hull of every blob in $\mathbf{B}$ is used to remove residues of the glint removal. In difference to [CE14], it is checked if the blob detection already detected the pupil. This is assumed to be true if the largest convex hull of each blob covers more than 70 % of the summed area of all blobs. Otherwise, blobs are merged whose center is closer

Algorithm 5 Detect occluded pupil $(\mathbf{I}, \mathbf{M}_L)$

$\mathbf{I} \leftarrow 2 \cdot \mathbf{I} - G * \mathbf{I}$
$\mathbf{I} \leftarrow F_{\text{Min}}(\mathbf{I}_{Filt}, k_{Min})$
$\bar{\mathbf{I}} \leftarrow \text{Normalize}(\mathbf{I})$
$\mathbf{M}_{PupilSeg} \leftarrow \left\{ p \in \Omega \mid \bar{\mathbf{I}}(p) > t_{Pupil} \right\}$
$\mathbf{B} \leftarrow \text{Blob Detection}(\mathbf{M}_{PupilSeg})$
$\mathbf{C} \leftarrow \text{Convex Hull Contour}(\mathbf{B})$
$\mathbf{C} \leftarrow F_{\text{Erode}}(\mathbf{C})$
$\left\{ p, e_x, e_y, \phi \right\} \leftarrow \text{Ellipse Fit}(\mathbf{C})$
return $\left\{ p, e_x, e_y, \phi \right\}$

to the center of the largest blob (Fig. 6.9c) than half the maximum extent of the largest blob d. The contour of the convex hull of the merged blobs then gives a first estimate of the pupil contour $\mathbf{C}$ (Fig. 6.9d). This contour is refined by first removing any point closer than 5 pixels to each other and then removing colinear points since those are probably generated by the (mostly) straight geometry of the eye lid (Fig. 6.9e). Finally, an ellipse is fitted to the remaining contour points to obtain position p, eccentricity e_x and e_y and angle ϕ of the projected pupil (Fig. 6.9f).

Partially Occluded Pupil The last case to treat is a strongly occluded pupil (Alg. 5). Here, the contrast is enhanced using unsharp masking; $\mathbf{I} = 2 \cdot \mathbf{I} - G * \mathbf{I}$ (Fig. 6.9h), where G is a Gaussian and $*$ is the convolution operator. A minimum filter with radius $k_{Min} = 21$ pixels is applied to remove eye lashes. By normalizing the input image $\mathbf{I}$ to the range $[0,1]$, pupil segments are detected by adaptive thresholding. Setting the threshold to $t_{Pupil} = 0.12 + (\|\widetilde{p} - p_E\|)^{0.5}$ yields an approximate mask of the pupil fragments where p_E is the pixel position of the center of the eye ball. This formula is used to compensate for an observed vignetting effect towards at the border of the eye.

As in Alg. 4, blob detection is performed and the resulting blobs are merged to estimate the convex hull of the result (Fig. 6.9h). To counteract the minimum filter, the result is eroded with a similar kernel of size k_{Min}. Finally, the contour of the blob is extracted and again ellipse fitting is performed to obtain the ellipse parameters of the projected pupil (Fig. 6.9j).

Fig. 6.10 Real-time Gaze-contingent Rendering. Foveated rendering (left): Rendering quality and color saturation is deliberately decreased towards peripheral vision. Adaptive depth-of-field effect (center, right): near and far focal distances. The gaze vector is shown as a red marker.

6.5 Applications

Several applications for the ETHMD with integrated eye tracking have been implemented which are based on the freely-available Unreal Engine and open-source game content [EPI15].

Adaptive Depth-of-Field Rendering Inspired by previous studies for desktop applications, a simulation of the accommodation reflex has been implemented [HLCC08, MCNV14]. In reality, accommodation allows us to focus on objects at arbitrary distances by flexing our eye lens (Chapter 2.4). In consequence, other objects appear naturally blurred. To compute the focal distance, a ray is cast into the scene starting at the viewpoint and directed along the viewing direction as measured by the eye tracker for one eye. Then, the distance is determined to the surface the ray hits first (Fig. 6.10), and the scene is rendered with the appropriate depth-of-field effect.

Foveated Rendering In the second application, it is shown that the gaze tracker enables simulation of a gaze-contingent display. Previous work showed the potential of foveated rendering techniques [RLMS03, DÇ07, GFD$^+$12]. Due to the rapid acuity fall-off from foveal to peripheral vision rendering can massively benefit from gaze contingency. The effect is demonstrated by rendering a scene at five different resolution levels corresponding to circles of different radii on the screen. The highest resolution corresponds to the foveal region where the user is looking at. The render resolution is reduced by a factor of two for each following circle. Between render resolution levels, pixels are smoothly blended to avoid visible resolution seams.

Fig. 6.11 Gaze Transfer and Avatar Animation. (Left, center) Eye tracking enables more expressive and natural character animation. The estimated pupil size and blink event can also be used to animate eye adaptation and blinks instantly (right).

The implementation shown in Fig. 6.10 is just simulating foveated rendering and does not lead to actual boost in performance. When using a path tracer, however, the number of samples per pixel can be reduced in the peripheral field of view, whereas for rasterization a lower level-of-detail or less texture lookups can be performed. A novel foveated rendering method using the proposed ETHMD is presented in Chapter 7. Additionally, similar techniques can be used to simulate various visual field defects, such as hemianopia (partial blindness), color blindness, retinitis pigmentosa (night blindness, blurring of vision, loss of central vision, and others) or pigmentary retinopathy (deposits of pigments, Fig. 6.10, left).

Gaze Transfer for Avatars In this application, immersion is enhanced by mapping gaze direction and head movements of the user onto an avatar standing virtually in front of him. The eyes of the avatar rotate as the user rotates his eyes and the avatar blinks as the user blinks (Fig. 6.11). This increases perceived realism for the user in VR and offers novel opportunities for self-expression. Gaze transfer can be a valuable extension in telepresence applications or user-to-user communication in the context of collaborative virtual reality. With multiple users wearing an ETHMD, immersion can be enhanced by enabling collaborating users to establish eye contact in VR.

Gaze Maps are an effective method to visualize the user's gaze over time and an effective tool for user experience studies [Sch14]. For a demonstration using the binocular eye tracker a stereoscopic movie player has been implemented. The software records gaze data while watching the video. Gaze maps for multiple users can be derived by plotting and filtering the estimated screen positions for all viewers. The result is shown for one frame of a movie in Fig. 6.12. Most viewers fixate the person in the foreground, the picture in the background or the table. A temperature color coding is used for visualization (hot areas are fixated more than cool areas).

Fig. 6.12 Gaze visualization. In a fixation map gaze directions are represented locally by circles. Circle diameters indicate the duration of each fixation (left). Gaze heat maps (right) show the fixated display area averaged over time or users. Temperature colors represent total fixation duration.

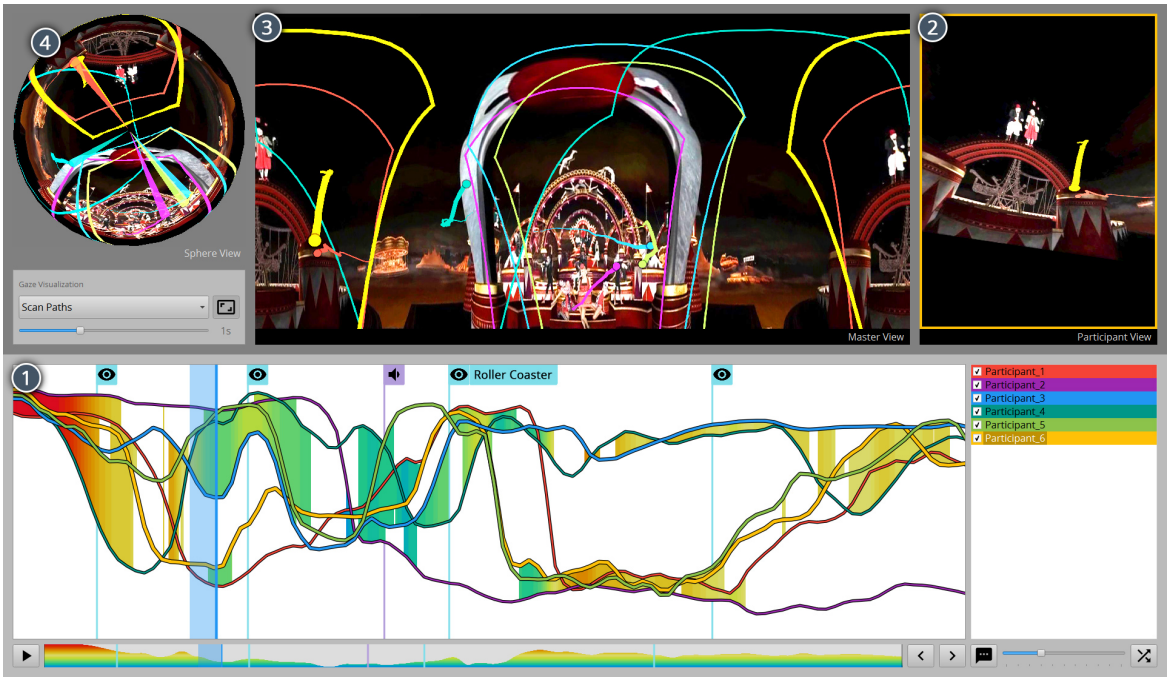

Fig. 6.13 Immersive gaze analysis. User interface to analyze viewing behavior of immersive live-action videos [LSF+15]: On top, color-coded frames indicate the current field of view of multiple users. Below, the users' scan paths allow analyzing gaze direction of all users over time. In immersive environments, gaze direction of different users diverge much more than for conventional TV.

VR Video Analysis and Rendering Being able to track gaze inside head-mounted displays is a necessary prerequisite for evaluating perception in immersive environments [LSF+15]. For example, 360° videos in HMDs are viewed fundamentally differently from conventional movies on a TV screen. Using the proposed ETHMD, a new visualization and analysis tools has been developed (Fig. 6.13) to investigate viewing behavior when immersed all around in live-action footage [LSF+15].

The notion of fovated rendering is also useful for broadcasting and for rendering and display of immersive 360° videos [DCM04]. Currently, video codecs are optimized for encoding blocks of pixels at the same resolution in every part of the video frame. In light of the retina's vastly varying perception characteristics from foveal to peripheral vision, however, future gaze-contingent video codecs are able to adapt coding rate to local view eccentricity. With gaze-contingent coding, only perceptually relevant information needs to be transmitted and rendered, saving bandwidth and memory. In terms of bandwidth and computational complexity gaze-contingent video rendering may be especially valuable for light field video playback and light field displays that render multiple viewpoints for each displayed frame.

6.6 Evaluation

In this section, the proposed method is evaluated by assessing tracking quality and performance. The pupil detection algorithm is tested against two other state-of-the-art algorithms [LWP05, CE14]. The section concludes with a user study with 33 participants.

Performance Evaluation The eye tracking framework has been implemented in C++ using the OpenCV algorithm library [OCV15]. The primarily CPU-based processing pipeline achieves a total end-to-end latency, from capturing the eyes by the cameras until a rendered frame is visible to the user, of 32 ms on current hardware (i7-4930K @ 3.4 Ghz, GeForce GTX 780 Ti). The pupil estimation of both eyes and the camera capture threads run in parallel on multiple cores of the CPU. Some of the pre-processing filters (sharpening, blur) run in CUDA on GPU. The eye-tracking camera resolution is 640×480 pixels at 75 frames per second. Timings for each step of the pipeline are as follows:

Process step	*Duration (milliseconds)*
Frame grabbing (@75 Hz)	≈ 13
Pupil estimation	≈ 9
Gaze estimation	< 1
Rendering (Application)	≈ 10
Total Latency	≈ 32

Table 6.1 End-to-end latency estimation.

Tracking Quality The pupil-tracking algorithm is evaluated in terms of tracking stability and tracking precision. After having calibrated the eye tracker for two different users with corrected-to-normal vision, the tracking precision was measured. The proposed algorithm is compared to the *STARBURST* eye tracking algorithm of Li et al. [LWP05] and the auto-threshold algorithm of Chen et al. [CE14]. Both algorithms work for near-field eye tracking without relying on corneal reflections. Glint-free images are used as input as required by these algorithms.

T1 Pupil Position and Size Pupil position and size is tested objectively against ground-truth data derived from manually created pupil masks of a 1987 frames video recorded with both eye tracking cameras. The error values for pupil position and size are computed by the differences of the extracted pupil-ellipse position and eccentricity. The result of the test is summarized in Table 6.2. The pupil position error ε_{Pos} is computed as the average pixel deviation of the computed position p_e from the reference position p_{gt}. In contrast, pupil size error is based on the eccentricities:

$$\varepsilon_{Pos} = \frac{1}{n}\sum_{i=1}^{n}(|p_e - p_{gt}|) \qquad \varepsilon_{Size} = \frac{1}{n}\sum_{i=1}^{n}(|e_x - e_{x,gt}| + |e_y - e_{y,gt}|) \tag{6.3}$$

where e_x, e_y are the eccentricities of the estimated ellipse.

In terms of accuracy, the novel algorithm clearly outperforms the competitors as they can not deal with partially occluded pupils. For each tested pupil detection algorithm, the pupil size is closest to the real pupil size for a central view. The pupil size artificially increases as the view tilts towards the sides due to increasing lens distortion, resulting in partial magnification of the projected pupil.

Test	ε_{Size} (px)	ε_{Pos} (px)
Ours	**0.04**	**2.16**
Auto-threshold [CE14]	0.63	21.67
Starburst [LWP05]	0.24	13.15

Table 6.2 Pupil position and pupil size accuracy.

T2 Gaze Direction Error An additional test evaluated the difference of the screen position returned by the eye tracker and the reference screen position set by a visible marker on screen as

$$\varepsilon_{Screen} = \frac{1}{n}\sum_{i=1}^{n}|s_e - s_{gt}| \qquad \varepsilon_{Ang} = \tan^{-1}\frac{\varepsilon_{Screen}}{d_{EyeScreen}} \tag{6.4}$$

where s_e and s_{gt} are the estimated and reference screen positions and n is the number of tracking samples (n = 30 in our test). The pixel error is transformed into the angular error by estimating $d_{EyeScreen}$ via ray tracing using the calibrated model. The error is evaluated for thirty different positions.

The error ranges from $\varepsilon_{Ang} \approx 0.5$ °to $\varepsilon_{Ang} \approx 3.5$ °, being generally higher at the borders of the screen due to stronger occlusion and therefore reduces pupil detection quality. The interpolated screen position error is visualized in Fig. 6.14.

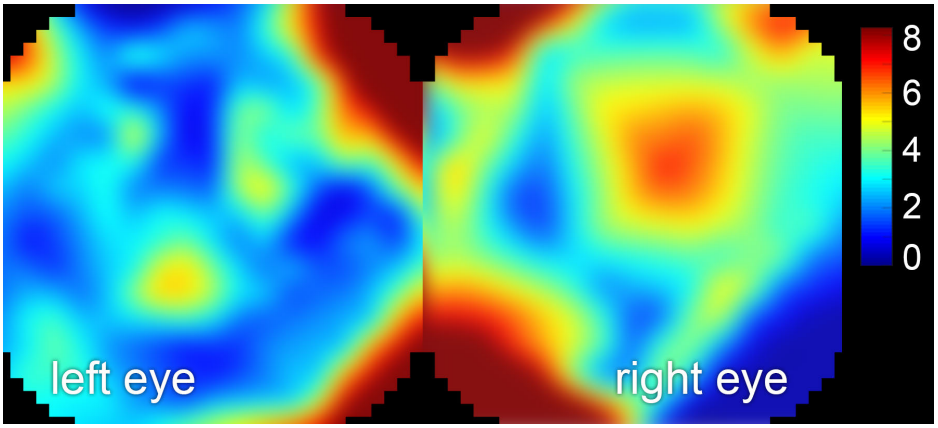

Fig. 6.14 Gaze direction error. The absolute error for both eyes over the available FOV, given in screen pixels.

User Study The ETHMD was tested by 33 participants (25 males, 8 females); 15 had normal vision, 18 had corrected-to-normal vision. The current prototype does not support wearing glasses when using the HMD. However, the lenses can be adjusted to compensate for a wide variety of ametropia and hypertropia [DFRR10]. Every person started with the user calibration procedure and then was able to use the adaptive DOF application (Sec. 6.5). Afterwards, the users were asked to rate certain aspects of the device (update rate, latency, stability, accuracy) and the application (naturalness, usefulness, user experience). The evaluation of the user feedback is visualized in Fig. 6.15. In summary, the user feedback was very positive with regard to user experience and usefulness of the system. Every user mentioned that they would prefer gaze use in many applications. The stability and accuracy was rated positive but not yet completely convincing. There were two major issues which explain the reduced rating. The system is currently an early prototype and still features disturbing redundant cables from the cameras, as well as an inflexible display cable, which resulted in slight shifts of the HMD when turning the head, and reduced the accuracy of the gaze estimation. Another issue for some participants was the usage of mascara on the eye lashes which negatively influenced the pupil estimation, resulting in a reduced user experience.

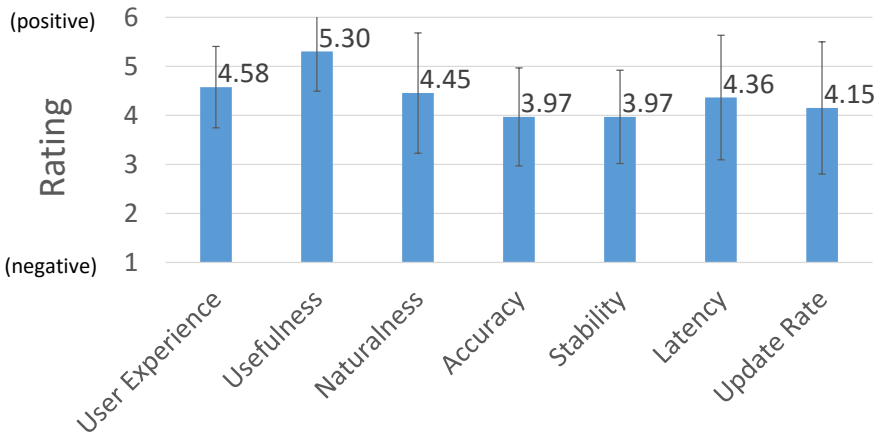

Fig. 6.15 User study results. Blue bars show user ratings concerning specific aspects. The scale ranges from 1 to 6 (negative/positive). Black bars represent standard deviation.

6.7 Discussion

Limitations The concept of adjustable lens holders provides sharp vision even for people usually wearing glasses. Wearing glasses inside the HMD is an open problem as this would require larger lens-to-eye distance, larger lenses, and a larger screen for the same field of view. A fixated positioning of the HMD with respect to the head is also crucial. Otherwise recalibration becomes necessary.

Long-term user study In this work, the tracking quality of the gaze estimation algorithm has been tested for only a small number of people and limited duration (several minutes) of usage. In the future, a larger user study should be conducted to improve the hardware design and software of our prototype. Additional studies with longer usage sessions will provide more information about robustness, usability and wearing comfort.

Auto-calibration In the literature software methods for auto-calibration are described that rely on natural scan pathes of the environment and provide a seamless transition between calibration and interaction phase [PVT+13]. These concepts may be beneficial for the proposed system. However, none of these methods have been tested within the ETHMD yet. Alternatively, by using additional hardware, it might be possible that the calibration process can be largely simplified or completely automated. Klefenz et al. and Kohlbecher et al. exploit the precalibration in a stereoscopic camera setup to track the pupil without additional user calibration [KHKH10, KBB+08]. Alternatively a depth sensor could provide valuable information about the actual anatomy of the individual eye. An automated calibration process seems interesting even if additional hardware would increase weight and complexity of the device.

Applications The presented applications only scratch the surface of possible VR scenarios. Many other applications are enabled with by user gaze, or at least could benefit from this input, e.g., gaze-based selection and manipulation, or studies on user interfaces.

Perception studies in VR simplify analyzing properties of human vision and attention. These insights may lead to methods that will improve viewing experience or accelerate rendering. Perception studies also enable evaluation of simulations in VR, e.g., in the field of assembly processes or training for aerospace, military or surgery, psychological therapy, or eye disease simulation.

Using eye tracking as an input device enables novel gaze-based interaction metaphors, hands-free interaction with Attentive User Interfaces (AUIs) or an additional communication channel. The user is able to express his interest naturally by gaze. With additional cameras the ETHMD prototype could be extended for Augmented Reality usage (AR) where hands-free interaction is beneficial and precise IPD estimation and calibration are very important. Instead of using a closed body, the mirror-based setup and gaze-estimation technique could also be used for See-Through HMDs.

6.8 Conclusion

In this chapter a complete binocular eye-tracking solution for head-mounted displays has been presented. The system relies on low-cost components that should be affordable for every user group. This aspect opens the door for a large variety of novel applications and contributes to progress in research. The prototype has been tested on a small group of subjects. For the future, a user study with a larger group of people could be conducted in order to improve pupil detection and user comfort. In addition, new ways for continuous and automatic user calibration could be investigated.

Chapter 7

Perceptual Sampling for Gaze-Contingent Real-time Rendering

Contents

Modern rasterization algorithms can generate photo-realistic images. The computational cost for creating such images is mainly governed by the cost induced by shading computations. With ever-increasing display resolution shading has become the limiting factor in real-time rendering, especially for wide field-of-view (FOV) displays such as head-mounted displays (HMD) or wide-screen projection systems. Perceptual graphics algorithms make use of characteristics of the human visual system (HVS) to render only what we can actually perceive to reduce shading computation time [RFWB07, GFD+12, HGF14, VST+14]. This chapter targets *gaze-contingent shading* (also known as *foveated rendering*), a subarea of perceptual graphics focusing on the exploitation of known gaze direction as estimated, e.g., by eye-tracking hardware.

7.1 Introduction

As visual acuity decreases in the periphery, computation time is wasted if the entire FOV is rendered at maximum resolution. Specifically, assuming a uniformly resolved pixel grid, at least 9k by 8k pixels are required to support foveal acuity over the full FOV for a person with normal vision. For next-gen HMDs companies aim for even higher pixel fill rates (16k by 16k at 240 Hz) to achieve aliasing-free, low-latency VR experience [RKA16]. In contrast, gaze-contingent rendering algorithms match rendered detail to what can actually be perceived by the user. With respect to shading, this can be achieved by using a ray tracing framework with *selective rendering* [CDdS06] by sampling densely in the central gaze area and more sparsely towards the periphery. In rasterization, this is more challenging due to the restrictions imposed by the rasterization pipeline. GPU hardware is optimized for rendering images at one constant overall resolution. Pioneering work towards *gaze-contingent rasterization* has shown that it is possible reduce rendering-time without affecting the perception of the displayed scene [GFD+12]. The idea is to render nested layers of increasing angular diameter and decreasing resolution which are blended to simulate a linear acuity fall-off. However, the resolution for each layer is constant and approximates the acuity very conservatively. In addition, simplistic linear acuity models do not model contrast sensitivity or other dynamic properties limiting visibility of stimuli.

In contrast, saliency methods have proven to be successful in modeling complex human perception properties [SC06, JDT12]. Most of the input required for traditional saliency estimation, however, is not known until shading takes place. Therefore, one of the remaining key challenges is how to adapt render quality to saliency, not just acuity fall-off, *before* actual shading.

To address this problem, areas of high attentiveness are estimated as part of a deferred shading pipeline right after the geometry pass and before actual shading. This combines the advantages of geometry-independent deferred shading with the reduced amount of necessary detail in non-foveal vision. Perceptual properties of the HVS are modeled to create a gaze-contingent sampling pattern. The model incorporates a selection of prominent cues such as gaze direction, visual acuity, eye motion, areas of high contrast, and brightness. This pattern is used to locally adapt shading computation and to shade only a small subset of image pixels while interpolating radiance among the remaining pixels.

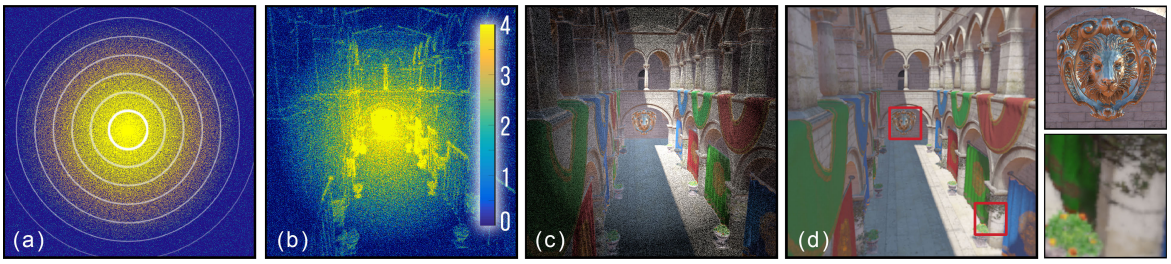

Fig. 7.1 Gaze-contingent Rendering Pipeline. Incorporating visual cues such as acuity (a), eye motion, adaptation and contrast a perceptually-adaptive sampling pattern is computed (b). Sparse shading (c) and image interpolation (d) achieve the same perceived quality as if shading each fragment, at a fraction of the original shading costs. The resulting image contains high object detail in the foveal region (lion statue inset) and reduced detail in the periphery (flowers inset).

The proposed approach is implemented into the eye-tracking head-mounted display (ETHMD) setup described in the previous chapter.
In particular, this chapter contributes:

- A flexible sampling scheme that is able to incorporate arbitrary perceptual cues (Sec. 7.3);
- An adaptive acuity model combining peripheral fall-off and eye motion (Sec. 7.3.1);
- A model for visual detail estimation in image-space combining spatial frequency adaptation for textures, perceptual filters in object- and screen-space (Sec. 7.3.2), and brightness adaptation (Sec. 7.3.3);
- A practical, smooth multi-rate rendering scheme for perceptually lossless gaze-contingent rendering suitable for a deferred shading pipeline (Sec. 7.4);
- A perceptual study validating the method (Sec. 7.5);

The gaze-aware shading method provides a general rendering algorithm for deferred shading. It is applicable to any type of display, reduces shading cost, scales sub-linearly with image resolution and FOV and leads to significantly reduced rendering times for sophisticated high-quality shading. Its flexibility allows incorporating any perceptual cue to control sampling of the shaded pixels.

The approach is related to Foveated 3D Graphics (F3D [GFD+12]), Multi-rate Shading (MRS [HGF14]) and Coarse Pixel Shading (CPS [VST+14]) (Chapter 3.2.6) but differs in several important aspects. F3D takes only acuity fall-off into account and renders the image with three layers of uniformly distributed samples of differing density. MRS and CPS divide the image into low-detail and high-detail shading, resulting in discrete shading rates which limit the effectivity for foveated rendering. In contrast, the approach described in this chapter provides a continuously decreasing sampling density fall-off. Additionally, the proposed adaptive approach models dynamic acuity changes over time induced by brightness adaptation, motion and higher-level object features. Therefore, the novel

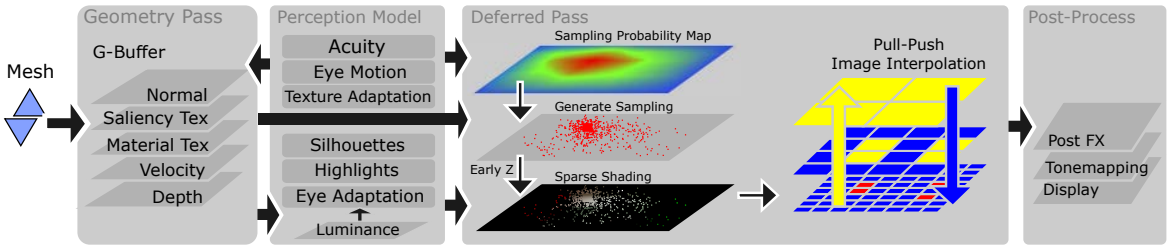

Fig. 7.2 Overview. The gaze-aware shading method is described in terms of a typical deferred shading pipeline. The geometry pass generates a G-Buffer at full resolution. G-Buffer data combined with predicted luminance, pre-computed object-based saliency and the visual perception model allow creating a sampling probability map. In the deferred pass, a sampling pattern is generated from the probability map which is then used for sparsely shading the image. Pre-processed material textures enable adapting spatial texture detail to visual acuity. A layered pull-push operation efficiently completes the missing parts of the image by interpolation. In the last step, post-processing operations like tone mapping and grading are applied before the final image is displayed.

technique can be seen as a generalization of F3D, MRS and CPS rendering suitable for foveated rendering and commodity graphics hardware.

7.2 Overview

The goal of the proposed rendering approach is to make use of characteristics of the HVS to determine and shade only the visually important pixels of a rasterized image. Quick interpolation of color values for the remaining pixels reduces shading cost and overall rendering time. The core of the approach consists of deriving a per-pixel probability function **P** for each frame, to decide which pixels should be shaded and which pixels can safely be interpolated. This process takes place *before* actual shading (Fig. 7.2). As available input information from the geometry pass is used, i.e., depth, normal, texture properties, etc. which is usually computed in modern rasterization pipelines such as Deferred Shading [ST90] or Forward+ [HMY13]. The algorithm assumes that the gaze direction **g** is known.

First, the features are described that are incorporated into the visual perception model for sample selection and how they are combined into the probability map **P** (Sec. 7.3). Then, it is described how the image synthesis step is implemented into a rasterization pipeline, including creation of **P**, sampling of **P**, shading of the selected samples, and interpolation of pixel color values for the final image (Sec. 7.4). To validate the effectiveness of the approach a perceptual study has been conducted (Sec. 7.5). Several experiments have been performed to analyze the efficiency of the algorithm (Sec. 7.6). In Section 7.7 strengths and weaknesses of the approach are discussed before the chapter concludes with Section 7.8.

7.3 Visual Perception Model

This section describes the extensible model to evaluate perceptual information content of an image to create the sample probability map **P**. The process is performed for each eye separately. Due to the real-time constraints, only quickly computable features are considered. All computations are performed in image space of the virtual camera before performing any optional image distortion for the particular output device, such as lens-matched warping compensating for the lens distortion of a VR headset. The goal is to derive a per-pixel sample probability map **P** that assigns a single *perceptual importance value* in the range $[0,1]$ to each output pixel based on a variety of quickly computable features $\mathbf{F}_0$ to $\mathbf{F}_n$. Higher values denote higher probability for correct per-pixel shading whereas lower values indicate an approximate shading may be used, e.g., fast color interpolation.

In the following, each feature $\mathbf{F}_i$ included in the model is described separately, distinguishing between acuity-based features (Sec. 7.3.1), attention-drawing features (Sec. 7.3.2), and global features (Sec. 7.3.3). Finally, the features are combined into a single sample probability map **P**.

7.3.1 Visual Acuity

Visual acuity provides an estimate of the smallest visual detail the HVS is able to resolve. In the model three related sub-features are considered: Acuity fall-off, eye motion, and brightness adaptation.

Acuity Fall-Off

Weymouth has postulated an approximately linear degradation behavior of acuity with eccentricity [Wey63]. Even though this model was proven to be valid only for eccentricities up to 30°, it is often used in methods for foveated rendering [GFD+12, VST+14, SMI16]. In the proposed approach the model of Weymouth is extended to make it better suitable for use with wide FOV. A constant acuity ω_p is assumed in the far periphery, as little is currently known about the acuity fall-off at eccentricities beyond 30° (Chapter 2.3).

The sampling probability $\mathbf{F}_\omega$ for a pixel position **p** and a gaze position **g** is then computed as:

$$\begin{aligned} \mathbf{F}_\omega(\mathbf{p},\mathbf{g}) &= \text{clamp}(f(\mathbf{p},\mathbf{g}),\omega_p,1)\ \text{, with} \\ f(\mathbf{p},\mathbf{g}) &= \omega_0 + m \cdot e(\mathbf{p},\mathbf{g})\ , \end{aligned} \tag{7.1}$$

where $\text{clamp}(f(\mathbf{p},\mathbf{g}),\omega_p,1)$ clamps the values of parameter f to the range $[\omega_p,1]$. The acuity limit ω_0 and acuity slope m are user-dependent properties which need to be set beforehand in a calibration step (Sec. 7.5). The function $e(\mathbf{p},\mathbf{g})$ computes eccentricity e in degrees.

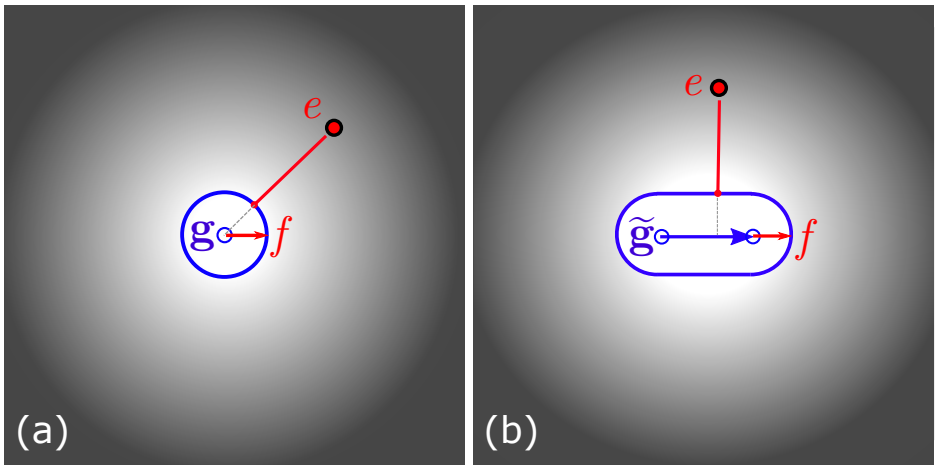

Fig. 7.3 Acuity-contingent sampling. Sampling probability is unity in the fovea (red arrow) and decreases with foveal distance (red line) towards the periphery. Equal foveal distances result in equal sampling probabilities. **Isotropic acuity fall-off (a):** During eye fixations the foveal region is defined by the gaze vector $\mathbf{g}$ and foveal radius f. Eccentricity is described by the foveal distance e. **Anisotropic acuity fall-off (b):** For smooth pursuit eye movement (blue arrow) the foveal region is linearly extended according to the gaze motion vector $\tilde{\mathbf{g}}$.

Eccentricity e is derived given the values for horizontal and vertical display resolution $\mathrm{d} = (\mathrm{w}, \mathrm{h})$, horizontal and vertical field of view $\mathrm{a} = (\mathrm{FOV}_h, \mathrm{FOV}_v)$, and values for gaze position $\mathbf{g}$ and pixel position $\mathbf{p}$:

$$\begin{aligned} \mathrm{s} &= (\mathrm{d}/2)/\tan(\mathrm{a}/2) \\ \mathrm{p} &= \mathrm{atan}(|\mathbf{p} - (\mathrm{d}/2)|/\mathrm{s}) \\ \mathrm{g} &= \mathrm{atan}(|\mathbf{g} - (\mathrm{d}/2)|/\mathrm{s}) \\ e &= \sqrt{(\mathrm{p}_x - \mathrm{g}_x)^2 + (\mathrm{p}_y - \mathrm{g}_y)^2}\ (180/\pi). \end{aligned} \tag{7.2}$$

An depiction of the resulting feature map is shown in Fig. 7.3a.

Next, eye motion is incorporated into the acuity model.

Eye Motion

Acuity $\mathbf{F}_\omega$ has been described for the assumption of a static gaze per frame, which is reasonable for high-refresh rate displays [GFD+12, SMI16]. However, smooth pursuit eye motion can invalidate this assumption for lower-refresh rates. To take eye motion into account two sub-features are incorporated: anisotropic scaling of the foveal region, and motion-dependent acuity adaptation.

Anisotropic Acuity Fall-off

To take the expected gaze motion into account during frame duration Δt, the gaze position is not modeled as a point but as a line $\tilde{\mathbf{g}} = \mathbf{g}_i + \Delta t \lambda \overline{\mathbf{g}_{i-1}\mathbf{g}_i},\ \lambda \in [0,1]$ where $\mathbf{g}_i$ and $\mathbf{g}_{i-1}$ are gaze positions during the current and previous frame. The foveal area and acuity fall-off are computed as described in Eq. (7.1) with the difference that in this case eccentricity e depends on the distance to a line instead of a position (Fig. 7.3b).

Motion-dependent Acuity Adaptation

Compensating for higher-latency displays requires increasing the foveal area, but sensitivity to detail in the HVS varies with respect to velocity across the retina [Kel79]. For example, in areas where the projected velocity of the displayed object deviates from eye motion in screen-space, the sample count can be reduced. Acuity computation is adapted from the work of Reddy on level-of-detail rendering [Red01]. Accordingly, the motion-dependent sampling probability $\mathbf{F}_M$ based on gaze motion $\tilde{\mathbf{g}}$ can be computed for any pixel $\mathbf{p}$ by:

$$\mathbf{F}_M(\mathbf{p}, \tilde{\mathbf{g}}) = \mathbf{F}_\omega(\mathbf{p}, \tilde{\mathbf{g}}) \cdot \mathbf{G}(\tilde{\mathbf{g}}/\Delta t)\ , \tag{7.3}$$

$$\mathbf{G}(\mathbf{v}) = \begin{cases} 1.0 & \text{if } \mathbf{v} \leq 0.83^\circ/s \\ 0.002 & \text{if } \mathbf{v} > 118^\circ/s \\ 0.962 - 0.463 \log_{10}(\mathbf{v}) & \text{else.} \end{cases} \tag{7.4}$$

7.3.2 Visual Detail

In this part features are introduced that may influence the user's attention and consequently should be taken into account during sampling. In addition, it is described how visual attention-drawing artifacts stemming from careless subsampling in the peripheral viewing area are avoided. Attention and gaze do not necessarily coincide. This is known as the concepts of foveal and attentional spotlights [Gol09]. Visual information is constantly processed in the periphery [SRJ11]. Different visual factors attract gaze more than others, specifically, regions of spatial and temporal contrast as well as saturated colors (Chapter 2.3). It is therefore important to faithfully represent these factors in the rendered image especially in the peripheral area, even though acuity is lower.

Because of this, a combined approach is used that

1. adapts spatial texture frequencies according to perceivable detail (Sec. 7.3.2),
2. extracts prominent scene geometry features based on a set of *perceptual filters* (Sec. 7.3.2).

This strategy is conservative in the way that, in order to avoid flickering from subsampling, it includes most of the existing high frequency details in the scene but reduces sampling probability in areas of visual indistinctiveness.

Texture Adaptation

Textures generally represent material properties and surface details which can potentially attract the user's attention. In Sec. 7.3.2 it is shown how to detect these details in object textures and how to adopt sampling probability in **P**.

In real-time rendering prefiltered textures, mostly mipmaps, are used to avoid aliasing by removing high frequencies in the textures based on the projected size of the texture in screen space. This approach is extended to incorporate also *resolvable detail* of the HVS, i.e. texture details projected in the peripheral area are removed. During mipmap creation each level is filtered using a Gaussian filter before subsampling the texture for the next level, thus removing higher frequencies contained per mipmap level.

During rendering the mipmap level is selected as follows: In screen space the projected texel size t_s and the corresponding solid angle t_{ang} in the user's view are computed. This value is then compared with the resolvable detail of the corresponding pixel encoded in the acuity function $\mathbf{F}_M(\mathbf{p},\tilde{\mathbf{g}})$. In case the acuity value is higher than the angular texel size the lowest mipmap level is proposed ($l = 0$). Otherwise, the mipmap level l is derived as follows:

$$l = \text{clamp}(\log_2(t_{ang}/\mathbf{F}_M(\mathbf{p},\tilde{\mathbf{g}})),\ 0,\ \#\text{mipmaplevel}). \tag{7.5}$$

This value is then compared with the traditionally computed mipmap level based on the projected size and the maximum of both is taken for the final texture lookup.

Perceptual Filters

In the following usage of three different filters is described to retrieve potential image regions of high contrast that may be perceptually significant according to previous saliency and psychophysical perception literature (Chapter 3.1.2). In the implementation part, computational aspects of each filter are explained (Sec. 7.4). The selection of detectors does not compromise a complete model for human perception simulation which is still an active research topic. Other perceptual as well as attentional cues affecting perception can be added easily to the perceptual model. The max-operation in Eq.7.11 allows selecting a fragment for rendering if any of the detectors gives a positive response whereas the scaling term allows including attributes that inhibit perceptual importance.

Object Saliency Detection Besides the usual object textures used for rendering, additional object saliency textures r_{Obj} at the same resolution are created for each mipmap level. Each texel in r_{Obj} corresponds to the perceptual importance of a point on the object's surface as follows:

$$r_{Obj} = \max\left(\bigtriangledown r_{Norm}, \bigtriangledown r_{Bump}, \bigtriangledown r_{Alb}, r_{Gloss}, r_{Met}\right). \tag{7.6}$$

For each of the normalized material parameters geometry normal r_{Norm}, detail normal r_{Bump}, and albedo r_{Alb} the maximum of the local partial derivatives is computed using a simple gradient filter of

kernel size 3×3. Shiny materials in physically-based rendering can easily draw the user's attention through highlights on the surface [WLC+03]. For this reason, the computed gradient values are compared with the normalized metalness r_{Met} and glossiness r_{Gloss} of the material. The maximum of the values gives the object saliency texel.

During the geometry pass the saliency textures are rendered into a separate buffer to form another feature map $\mathbf{F}_O$.

Silhouette Detection View-specific regions of potential contrast cannot be precomputed. Therefore, a silhouette detection filter is employed that works on scene geometry from the viewpoint of the observer. The detector is essentially a gradient filter responding to changes in scene depth d and normal direction n. Both channels are combined by taking the maximum to form the silhouette feature map $\mathbf{F}_S$.

Highlight Detection Last but not least, a highlight detector is used since our eye is sensitive to regions of high contrast [SRJ11]. The detector works as described by He et al. [HGF14]. This detector enforces higher sampling probability for pixels that could potentially contain bright highlights, which, if not sampled properly, could otherwise lead to flickering artifacts. The detector is computed as

$$\mathbf{F}_H(\mathbf{p}) = \sum_i \langle \mathbf{L_i}(\mathbf{p}), \mathbf{R}(\mathbf{p}) \rangle^\gamma \cdot \mathbf{I_i} \ , \tag{7.7}$$

where the dot product of the normalized light direction vector $\mathbf{L_i}$ and the normalized reflection vector $\mathbf{R}$ is scaled by the intensity $\mathbf{I_i}$ of the i-th light source. Contrast is increased using a power function with a large exponent ($\gamma = 20$). The result is clamped to zero to avoid negative light influence. Computing a dot product for each light source is inexpensive compared to full shading of the image.

7.3.3 Brightness Adaptation

Adaptation is the time-dependent process when the eye slowly *adjusts* to the surrounding lighting situation (*cf.* Chapter 2.3). Visual acuity for details and color perception is reduced in low light conditions. On the contrary, during daytime sharp vision and color vision work very well [LSC04, MDK08, EJGAC+15]. In addition, glare reduces the ability to see clearly, i.e., if light of a very bright light source enters the eye [EJGAC+15].

In the proposed approach the idea of global luminance maps [PY02] is used to adjust the sampling probability according to eye adaptation. As luminance information is not available prior to shading, the fact is exploited that adaptation is no instantaneous effect but a process over time. Therefore, a low-frequency luminance map from the previous shaded frame I_{i-1} is used. Based on the RGB values before tone mapping the luminance map is computed as:

$$L(\mathbf{p}) = \langle (0.299, 0.587, 0.114)^\top, I(\mathbf{p})_{i-1} \rangle \ . \tag{7.8}$$

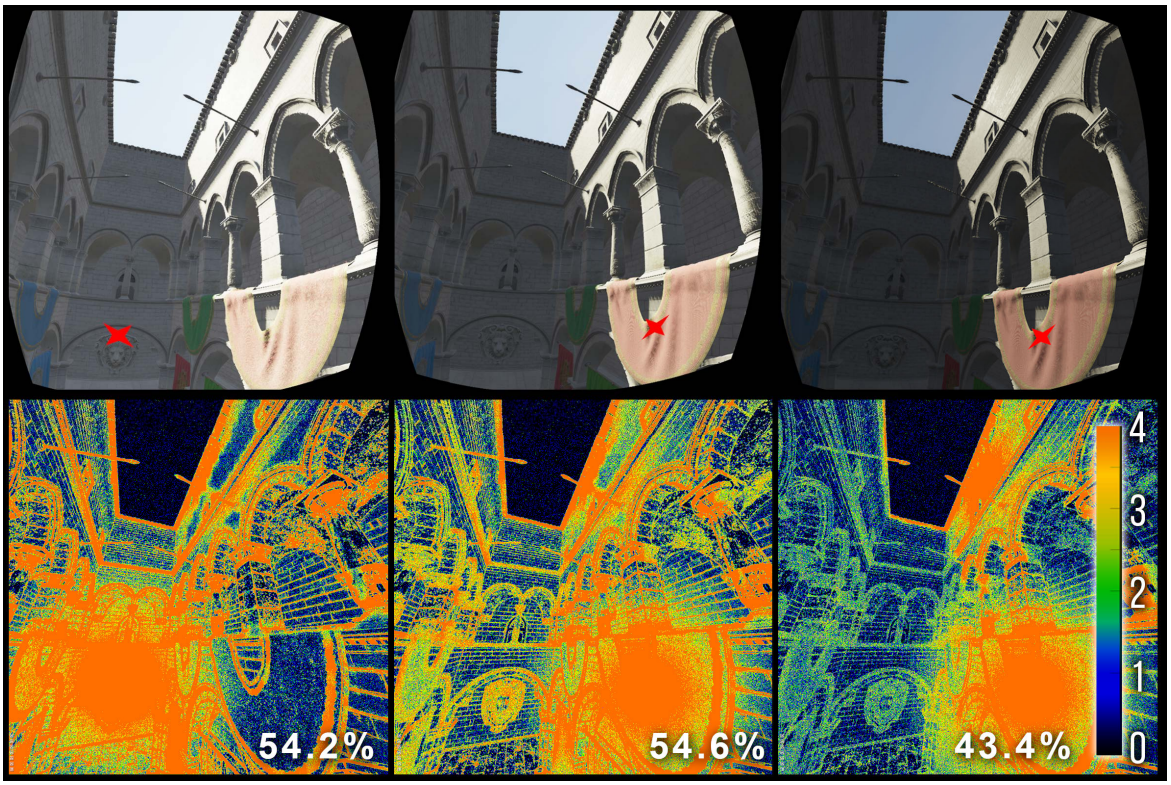

Fig. 7.4 Brightness-adaptive sampling. The sampling scheme distributes samples in the periphery in accordance with time-dependent adaptation. As over-exposed (left, bright wall) and under-exposed regions (right, shadow area) contain less perceivable details compared to normal exposure (center) sampling probability is reduced. The relative shading count vs. per-pixel reference is given in percent. The color-coding represents the number of shaded pixels in a 2×2 neighborhood (see legend).

Exposure is iteratively updated for each frame using an empirically derived adaptation rate $a_r = 0.05$ as

$$\begin{aligned} E_i &= E_{i-1} + a_r \cdot (A - L_{avg}) \text{ , with} \\ L_{avg} &= \tfrac{1}{N} \left(\textstyle\sum_{\mathbf{p}} L(\mathbf{p}) \cdot \mathbf{F}_M(\mathbf{p}, \tilde{\mathbf{g}}) \right) \text{ ,} \end{aligned} \tag{7.9}$$

where E_i is the new exposure value, E_{i-1} is the previous frame's exposure, A is the user-defined auto exposure value and N is the number of pixels. Finally, I_{i-1} is tone-mapped based on the adjusted exposure E_i [RSSF02] and converted again into a luminance map $\bar{L}$ using Equation (7.8) .

Based on the tone-mapped expected luminance distribution $\bar{L}$, a scaling function $\mathbf{S}$ is computed for the per-pixel sample distribution $\mathbf{P}$ using the following equation:

$$\mathbf{S}(\mathbf{p}) = \min\left(1, \frac{\bar{L}(\mathbf{p})}{t_{\text{dark}}}\right) \cdot \min\left(1, \frac{1 - \bar{L}(\mathbf{p})}{1 - t_{\text{bright}}}\right) \tag{7.10}$$

Using empirically estimated values for $t_{\text{dark}} = 0.15$ and $t_{\text{bright}} = 0.9$, the scaling function $\mathbf{S}$ linearly scales the darkest and brightest pixels, reducing sampling probability in under-exposed and over-exposed parts of the image (Fig. 7.4).

Finally, the per-pixel sample probability map is computed as the scaled maximum of all features:

$$\mathbf{P}(\mathbf{p}) = \max(\mathbf{F}_M(\mathbf{p}, \tilde{\mathbf{g}}), \mathbf{F}_O(\mathbf{p}), \mathbf{F}_H(\mathbf{p}), \mathbf{F}_S(\mathbf{p})) \cdot s(\mathbf{p}) \tag{7.11}$$

Hence, the probability of sampling a pixel is set according to the highest importance value returned by any of the feature maps which assures that visually important pixels are not missed.

7.4 Implementation Details

Decreasing the number of shading samples is most beneficial for high-quality render methods. For this reason the proposed technique is implemented in a deferred rendering pipeline as used by most AAA game titles [KG13]. Applicable effects include physically-based rendering [CT82, Sch94], many-lights methods, image-based HDR environment lighting, screen-space reflections [SNRS12], HDR bloom effects, and adaptive tone mapping. An overview of the pipeline is given in Fig. 7.2.

The saliency textures are computed offline (Sec. 7.3.2). Rendering starts by computation of the motion-compensated acuity feature map $\mathbf{F}_M$ from gaze position and gaze motion as provided by the eye tracker. The acuity feature map is used as input to the subsequent geometry pass.

Geometry Pass In the geometry pass the scene is rendered from the camera view and rasterized into the G-Buffer. The G-Buffer consists of channels for world position, normal, depth, velocity, as well as material data such as albedo, roughness, metalness, and cavity. For the precomputed object saliency (Sec.7.3.2) the G-Buffer has one additional 8-bit saliency channel. In addition, apart from image-based lighting shadow maps are computed for the active light sources.

Perceptual Probability Density Function In a second pass the perceptual filters are applied (Sec. 7.3.2) to the information available in the G-Buffer. Computation results of each filter are gathered in the single-channel sampling probability map. The silhouette features detector checks for virtual edges in the scene (Fig. 7.5b) based on the normal and depth map of the G-Buffer (Sec. 7.3.2). Although texture normals have already been analyzed for each object in a preprocessing step, there may be regions of contrast due to object penetration or at silhouette boundaries. Next, object-based saliency is included (Sec. 7.3.2). Since the saliency texture value depends on motion-based acuity when being written to the G-Buffer, texture values from the appropriate mipmap are used as is (Fig. 7.5c). Then, highlight detection is applied for each visible light in the scene (Fig. 7.5d) by rendering light meshes, evaluating the detector in the fragment shader for all affected pixels, and accumulating the results (Sec. 7.3.2). For directional lights the filter is applied to the full image by rendering a screen-aligned quad. The brightness adaptation scaling function (Eq. 7.10) is estimated based on a

low-resolution version of the previous frame (same resolution as the acuity function texture). The feature maps are combined according to Eq. (7.11). Extending the feature map $\mathbf{F}$ with a 5×5 dilation kernel sufficiently enlarges salient feature regions to avoid potential flickering in shading. To speed up this costly process a mipmap representation of the feature map is used. Dilation is computed with a small kernel on a lower-resolution version of the feature map $\mathbf{F}$. Finally, we use the combined probability map $\mathbf{P}$ to generate the importance sampling pattern used for shading (Fig. 7.5e) [Rey98]. For every pixel $\mathbf{p}$ we compute a pseudo-random number p and create a shading sample if $p < \mathbf{P}(\mathbf{p})$.

Deferred Shading Pass and Sample Interpolation Instead of saving the sampling pattern explicitly in a binary texture, the largest depth values are reserved in the depth buffer to encode the sampling pattern and to scale the other depth values accordingly. Rendering a screen-filling quad using early z-culling then invokes the fragment shader only for the shading samples while efficiently discarding the rest, which is faster than discarding individual pixels in the shader. Limitations of this approach are discussed in Sec.7.6.

Shading from all primary scene lights is accumulated first. Then, the computed radiance values are interpolated as described in the following. Computation of secondary lighting effects in screen space, such as reflections or global illumination, are postponed up until after interpolation.

Image interpolation is possible due to the assumption that every perceptually important detail is included in the sampling. Remaining parts can therefore be interpolated without being noticeable by the observer (Fig. 7.5f-g).

Edge-aware interpolation would be too costly for this step. Even the Guided Image Filter [BEM11] – a very efficient version of a bilateral filter – takes too long. Instead, a fast GPU version of pull-push is used for interpolation [GGSC96]. This algorithm is based on mipmaps to fill in the missing shading information for all pixels and requires only four texture lookups per pixel in total. The generated mipmap levels can be reused in secondary shading effects, e.g. screen-space reflections, HDR bloom effects as well as for adaptive tone mapping. Those techniques require averaged or blurred radiance values, anyway [KG13, SNRS12]. Therefore, the radiance interpolation technique does not introduce significant overhead to the render budget.

In a final step of the post processing pipeline, gaze-contingent, temporally-adaptive tone mapping is performed based on the averaged color information generated by the pull-push step and the updated exposure, value as described earlier.

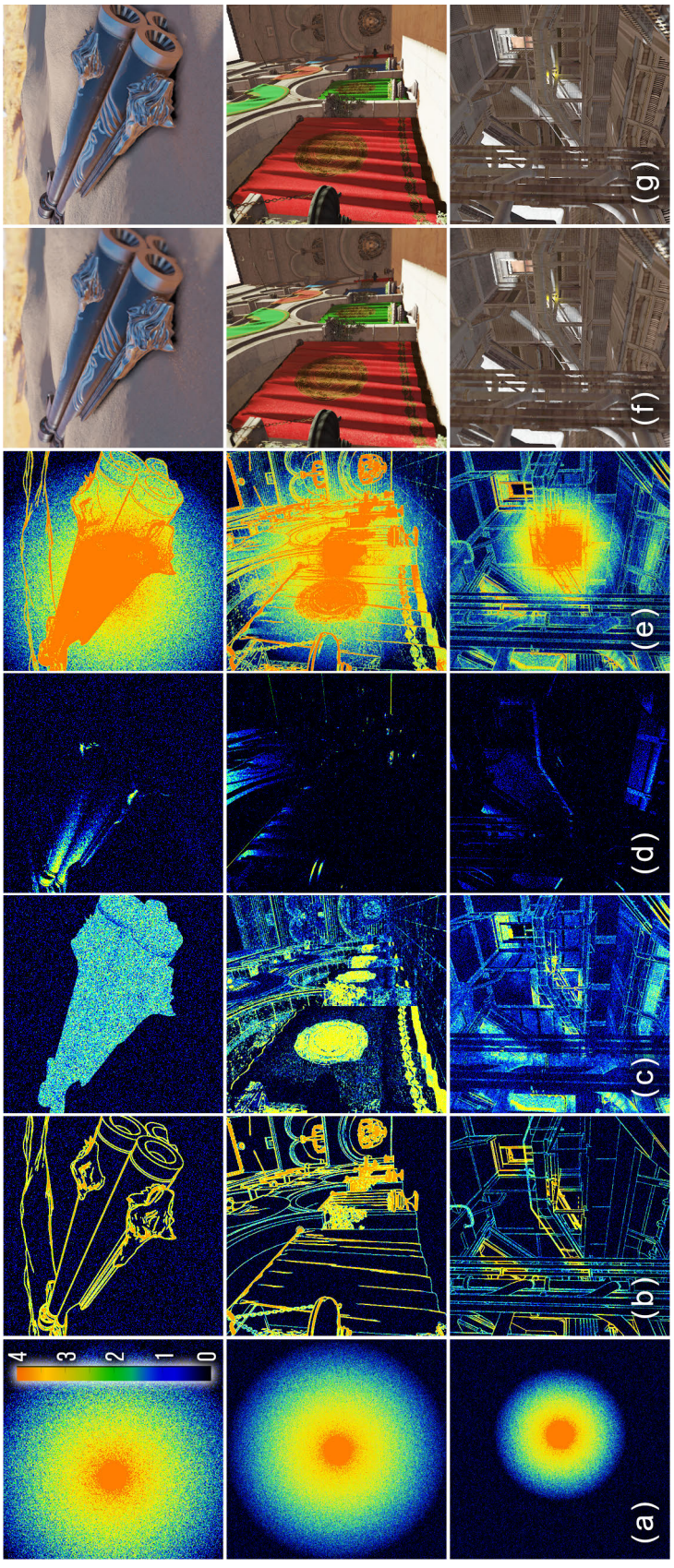

Fig. 7.5 Sampling and shading results for three different scenes. Our renderer includes different visual cues such as visual acuity (a), silhouettes (b), object saliency (c), and specular highlights (d). The combined features allow perceptual sampling (e) for sparse shading. Pull-push interpolation fills in missing pixels resulting in a complete image (f) which is perceptually equivalent to the per-pixel shaded reference (g). Color-coding in (e) represents the number of samples in a 2x2 neighborhood.

7.5 Perceptual Study

To validate that the results of the gaze-aware shading algorithm are visually equivalent to an image rendered with full per-pixel shading, a perceptual study has been conducted. The test scenes are rendered on a common desktop computer with an i7-4930K CPU and NVIDIA GTX 780 Ti graphics card with 3GB of GPU memory and displayed on the binocular eye-tracking head-mounted display (ETHMD) presented in Chapter 6. Eye tracking and rendering are performed in parallel. The measured latency of the configured eye tracker is 12.5 ms. A worst-case latency of ≈50ms may happen right after saccading eye motion and before the system can adjust itself correctly again. This delay is tolerable in this case as blur detection of the HVS does not increase significantly for up to 60 ms due to postsaccadic suppression [LW07].

7.5.1 Acuity Calibration Study

As acuity fall-off is a user-dependent property, a calibration study has been conducted first to conservatively find well-working parameters for the size of the foveal region f, acuity limit ω_0, acuity fall-off m and minimal acuity in the wide periphery ω_p (Eq. 7.1) and to avoid time-consuming calibration later on. The algorithm was explained to six participants. Their task was to conservatively adjust the mentioned three parameters until they did not perceive any visual difference between a full rendering and our gaze-contingent rendering between which they could toggle at will. Three different test scenes have been presented in the test (Fig. 7.5). For the ETHMD setup the following parameters (Eq. 7.1) were found: The average acuity limit in normalized device space was $\omega_0 = 1.1715$. The linear acuity slope is $m = -2.45$. The average size of the per-pixel shaded foveal region is $f = 0.07$. For the wide periphery ($e > 0.47$) the minimal acuity was found to be $\omega_p = 0.02$. The estimated conservative parameters are then used in the second experiment with a larger group of users which is described in the next section.

7.5.2 Validation Study

In this study the perceived quality of gaze-contingent rendering has been evaluated by assigning different visual tasks to the participants. The study was conducted with 16 persons (13 males, 3 females with corrected-to-normal vision) who had not used the system before and had not been informed about the strategy of our method. The following tests were performed targeting different aspects of the proposed technique. The user had to fulfill given tasks in a virtual environment (Sponza). Six trials were performed per test in which the respective test parameter was randomly activated or deactivated. For better comparison after each trial the screen switched to gray with a marker to focus the user's view on the screen center again. After each trial pair the participant was asked for perceived visual quality differences by choosing between the options "first better", "equal" or "second better".

T1 Cognitive load. The goal of this test was to draw the attention of the user to a specific and comparable task ("Count the colored spheres in the environment."). The positions of the visible spheres forced the user to look around and rotate his head in the virtual environment. The camera position automatically moved forward in the scene on a pre-defined camera track over 20 seconds. In each trial the colors and positions of the spheres changed randomly. Each trial randomly activated either the proposed gaze-aware method or the per-pixel shaded reference.

T2 Free viewing. In this test the task was to freely explore the environment without having a specific task to make it easier to detect quality differences. The time was constrained to 8 seconds. Again, each trial activated randomly either the gaze-aware method or ground-truth reference.

T3 Toggle manually. In this test, like in the calibration study, the user was able to toggle manually between our sampling and the reference as often as desired. This test was performed without time constraints. Therefore, the test was conducted only once per person.

T4 Brightness adaptation. In this test the user was asked to test the eye adaptation feature. In the virtual environment the user was seated in front of a wall partly lit by the sun, leading to under-exposed shadow areas or over-exposed lit parts depending on the user's gaze. The adaptive sample reduction for over- and under-exposed image regions was randomly activated or deactivated in each trial. As before, the user was asked for visual quality differences.

T5 Eye motion. In this test the eye motion-based sampling was activated/deactivated randomly to examine if the user perceives the reduced amount of detail in image parts moving differently to gaze motion. The user was asked to focus for 8 seconds on a sphere moving into the virtual environment. The moving sphere allowed to trigger smooth pursuit eye motion that is repeatable for each trial.

T6 Texture adaptation. In the last test the texture adaptation feature was randomly switched on/off while the user was freely exploring the environment. The goal of this test was to validate that the acuity-based peripheral textural detail reduction is not perceivable by the user.

7.6 Results

7.6.1 Shading Costs

Shading Samples The algorithm has been tested using three scenes featuring different characteristics. The *Cerberus* scene (Fig. 7.5, first row) shows an old revolver in the sand which contains many specular highlights due to the shininess of the metal. The *Crytek Atrium* scene (Fig. 7.5, second row) contains complex illumination including high contrast between shadows and lit areas as well as shadow edges of different intensities. The *Mothership* scene shows the engine room of a giant spaceship containing a high amount of geometric detail (Fig. 7.5, third row).

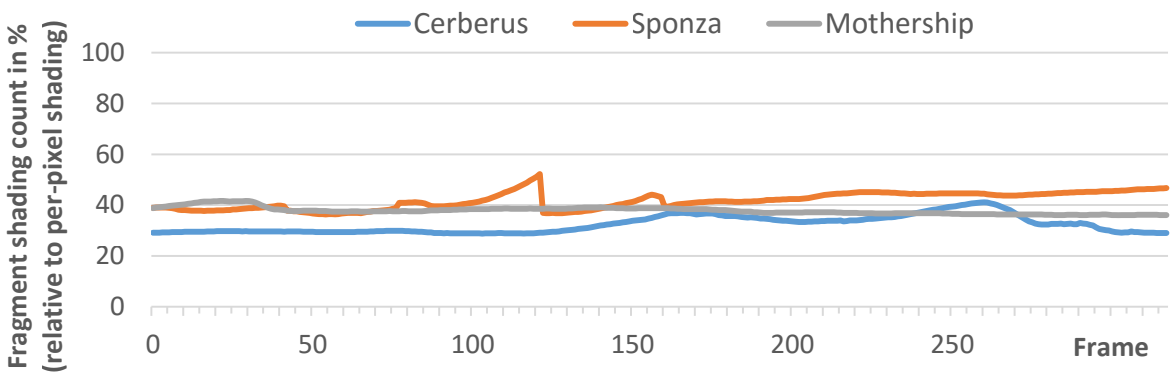

Fig. 7.6 Benchmark results. The amount of evoked fragment shader calls is shown for three scenes of 300 frames each. The adaptive method achieved *comparable* shading benefits in relation to the fully-shaded reference and *temporally stable* sampling rates for each test scene.

In all cases the amount of fully shaded pixels was reduced to roughly one third of the fragments (32.3% (Cerberus), 41.1% (Sponza), 37.2% (Mothership)). A detailed analysis over 300 frames is given in Fig. 7.6.

Render time The computational overhead of the proposed method is very low, about 0.9ms for the sampling creation step for both eyes (see Sec. 7.5 for PC specifications and resolution), and about 1ms for the shading interpolation, which is similar to creating mipmaps. As these are needed for many effects, anyway (depth-of-field, adaptive tone mapping, rough screen-space reflections, etc.) image interpolation comes almost for free. Using the implemented non-optimized renderer prototype, an overall reduction of rendertime of 25.4%, on average, has been achieved for the tested scenes. Full rendering for both eyes requires 15.1ms per frame, in average for all scenes, whereas the proposed approach requires only 11.2ms. Shading time is reduced by 41% (11.9ms to 7.0ms). The performance gain using adaptive sampling generally depends on the shading cost per pixel. Benefits are more significant if per-pixel shading is expensive. The difference between shaded pixels and savings in render time is mostly due to a hardware feature of current GPU architectures which always shade a 2×2 fragment group concurrently, no matter how many fragments are discarded [KNC16]. This limitation increases the amount of shaded pixels from 37 to 59%, on average, and therefore significantly reduces the shading benefit of the method on current hardware. Hopefully, next-generation GPU hardware will remove this constraint. Integrating hardware constraints directly into the sampling procedure has not been investigated yet.

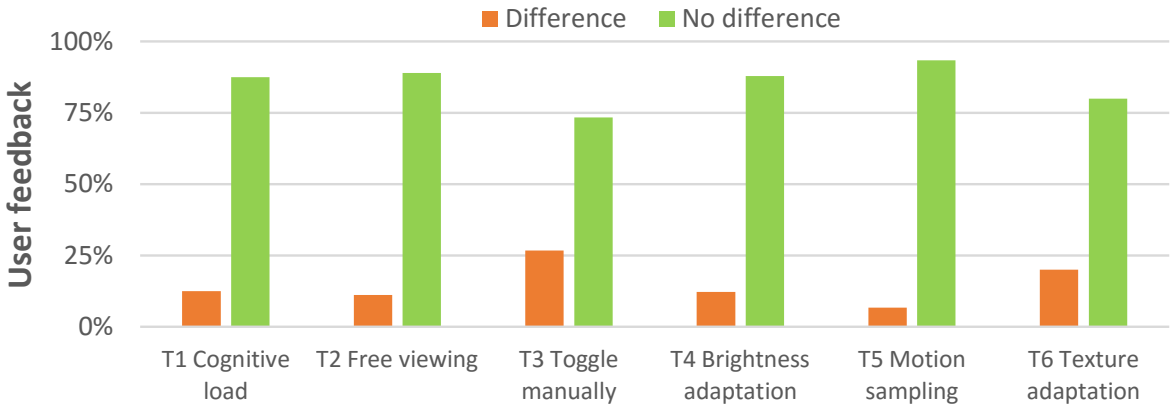

Fig. 7.7 Perceptual study results. Green bars show user ratings certifying visual equivalence between the perceptual sampling method and the full-resolution rendering in a variety of test scenarios. Orange bars show user ratings favoring the latter.

7.6.2 Perceptual Study Results

For the perceptual study a negative rating was given if the user *correctly recognized* the gaze-adaptive image (Fig. 7.7, orange). In the other case the user has either seen no difference or rated the per-pixel shaded reference as being worse (Fig. 7.7, green). In the passive viewing tasks (T2, T4, T6) most of the users have not detected a difference between the rendering variants. This validates the quality of the proposed perceptual model using the conservative acuity parameters derived initially. Importantly, the results of T1 and T2 don't expose any discernible difference: 87.5% (T1) and 88.9% (T2) of the users perceived no difference. Hence, the perceptual sampling performs equally well regardless of the user's cognitive load. Even in the most demanding case, T3, when comparing sampling and reference directly, the users have been able to recognize any difference in the images only in 26.6% of the trials but phrased the differences as "unobtrusive". Some users who perceived a difference had no preference for choosing the better-looking variant. The same holds for T6, testing texture adaptation to the acuity limit. Some users reported perceiving the changed amount of detail in the periphery (20.2%). In these cases the acuity slope could be increased resulting in more texture detail and equivalently more shaded samples. Although not being the intention of this work, some people actually preferred the look of the "smoother-looking" image. It may be reasoned that people perceived a reduction of aliasing in the periphery which may occur in the per-pixel shaded reference. User feedback has been very positive in T4 since the users liked the natural behavior of foveated brightness adaptation. The additional reduction of samples in washed-out image regions rarely received a negative rating (12.2%). In T5, hardly any user (6.7% of the trials) recognized a difference between sampling the acuity fall-off only and sampling including motion-based reduction. Interestingly, the motion-based sampling reduction often reduces the number of shaded pixels to just 10 to 15% when the eye tracks moving objects or when the user moves in the environment.

7.7 Discussion

In the following advantages and drawbacks of the proposed pipeline are discussed and compared with related approaches.

Applicability Many modern game engines use tiled deferred shading. Lately, also the Forward+ method was proven to be very efficient for many-light scenes [HMY13]. The perceptual sampling method can be applied to both approaches since both provide depth information before shading and support early z-discard to avoid expensive shading invocations. The GPU implementation could also apply to mobile hardware with gaze-tracking; the GearVR^{TM} has already been tested in combination with eye-tracking [SMI16].

The sampling technique scales favorably with resolution and FOV which are important properties for next-generation HMDs. The number of shading samples increases sublinearly with FOV whereas image interpolation scales linearly with resolution. In the tested scenes the shading samples decrease from 35% on 1.2k resolution to 17% when frame resolution is doubled (2.5k per eye). The adaptive approach provides temporal stability and predictable performance, as shown in Fig. 7.6.

Memory Consumption The required object saliency textures correlate to the overall texture usage in a scene. The sampling uses parts of the already available depth buffer. Other than that, the memory consumption is essentially equal to standard deferred shading. The mipmap required for the interpolation is created in any case for effects like screen-space glossy reflection etc.

Anti-Aliasing Anti-aliasing techniques reduce flickering caused by undersampling of geometric and shading details. Hardware-based multisample anti-aliasing reduces the causes of geometric aliasing but cannot handle aliasing stemming from undersampling highly specular materials, both spatially and temporally. Temporal anti-aliasing strategies, such as TXAA [KG14], solve this by accumulating shading information along pixel trajectories over time.

Although not yet implemented in the renderer prototype, the perceptual sampling does not prohibit usage of TXAA. In the foveal region each pixel is shaded so that TXAA can be applied as usual. In the periphery, flickering specular highlights are avoided due to texture adaptation (Sec. 7.3.2) which reduces frequencies in every material channel. Pixels with specular highlights induced by small-scale geometry are sampled due to the highlight detector. In this respect, TXAA should eliminate flickering specular highlights also in the peripheral viewing areas.

In the conducted study one user perceived peripheral flickering stemming from under-sampled shadow edges, which are not explicitly handled by our model but could be included by the extension of He et al. [HGF14]. However, modern shadow algorithms are optimized to produce appealing soft shadows, reducing this artifact by default.

Comparison to prior work In the following we compare the proposed approach to the related work in foveated 3D graphics (F3D) [GFD+12], multi-rate shading (MRS) [HGF14] and coarse pixel shading (CPS) [VST+14]. Both MRS and CPS are currently only theoretical concepts, implemented in software simulators. Both require adaptive shading features which are not available on commodity hardware. Only F3D and the presented perceptual sampling are directly applicable to current GPUs. Looking at the shading rates from MRS, CPS and adaptive sampling, comparable numbers are reported with relative instruction counts of about 30 - 70 %, depending on the scene complexity. This is reasonable since all mentioned approaches shade accurately in regions of high contrast and lower shading quality in low-contrast regions. F3D reports shading reductions by a factor of 10–15. Several reasons cause this discrepancy:

F3D theoretically undersamples the image drastically which results in very high frame rates. To diminish the resulting visible aliasing artifacts, F3D needs to rely on specifically designed anti-aliasing strategies, including jittered sampling of the image plane, temporal reprojection, and high-refresh rates to make use of eye integration over several frames. However, this limits F3D to simpler material models and less complex geometry. In contrast, adaptive sampling concentrates on carefully selecting samples for each single frame which allows incorporating additional features into the visual perception model besides acuity. In addition, the requirements of low-latency eye tracking hardware and high image refresh rates are relaxed by explicitly incorporating eye motion into the sampling model and accurately selecting samples around salient geometry and material features. This increases shading sample count but is necessity to be less dependent on specific anti-aliasing techniques or certain refresh rates.

A novel feature introduced in this work, from which F3D, MRS, and CPS can benefit, is texture adaptation. The efficiency of MRS and CPS heavily depends on scene content and material complexity. By reducing texture detail to the visual capabilities of the viewer, the proposed method is effective for arbitrary materials, including bump mapping.

MRS and CPS use a regular grid at three different shading levels. F3D also uses *discrete* resolution layers. In contrast, the novel sampling method may vary image quality *continuously*, giving the underlying model of the HVS a much higher level of flexibility. Considering the above similarities and differences, the perceptual sampling technique can be seen as a generalization of F3D, MRS and CPS rendering that is suitable for gaze-contingent rendering on commodity graphics hardware.

Future Work Further performance improvements may be achieved by collecting material data late in the deferred pass (deferred texturing). Then texture look-ups are only executed for those pixels that are actually shaded. In this case, the G-Buffer material data reduces to a material ID and UV coordinates instead of holding all material data. A straight-forward idea to improve performance for gaze-contingent rendering is view-dependent geometric level-of-detail [Red01]. Related approaches rely on drawing less vertices in the periphery or on reducing tessellation [WLC+03]. However, the HVS reacts quite sensitive to geometry changes. An analysis of a combined method of geometry and shading adaptation rate is a promising research direction. There is also promising work on decoupled

sampling that renders defocus and motion blur with less samples by introducing a memorization cache for reusing samples across visibility samples [CTM13, RKLC+11]. Mauderer et al. show that accommodation simulation has a positive effect on depth perception also for HMDs [MCNV14]. However, gaining performance benefits from accommodation in real-time rendering is still an open problem.

7.8 Conclusion

In this chapter a novel rendering paradigm for gaze-contingent rendering has been presented which combines the benefits of sampling flexibility and fast rendering based on a deferred shading rasterization pipeline. Previously, the flexibility to adjust shading quality accordingly to perceptual properties has been only available in ray-tracing approaches. Adaptive image-space sampling creates images that are *perceptually equal* to images rendered with full per-pixel shading, but at significantly reduced shading costs. For typical images of 1.2k resolution per eye, the method selects only about 30-40% of the pixels for shading while interpolating the rest. For higher resolution and wider fields-of-view, e.g. for HMDs, the amount reduces down to 20% of the original number of pixels.

The approach is universally applicable in the sense that perceptual sampling and rendering can be adapted to a variety of models to describe what attracts a user's attention and, what is equally important, what can be computed *before* the actual shading takes place. While the approach reduces shading cost, it does not reduce the cost for the geometry pass which is rendered at full resolution. This is necessary in order to predict the visually important parts of the image. Otherwise, it would not be possible to reliably and robustly detect silhouettes or fine surface detail before actual shading.

In the future, further refinements of the visual perception model should be investigated. Currently, the human vision model is tuned rather conservatively as the number of samples is potentially overestimated in order to not miss any attractors. It would, however, be interesting to see if one can compute the *minimal* number and positions of required samples. As this is partially user-dependent, further research is required, in graphics as well as in the field of psychophysics.

Chapter 8

Conclusion and Future Directions

Contents

8.1 Conclusion

Knowing where we are looking allows gaze-contingent display algorithms to provide for a much enhanced viewing experience. Whether by avoiding aliasing artifacts, by providing superior perceived visual quality or by allocating computational resources more efficiently, gaze-contingent computational methods are able to boost visual fidelity for all kinds of conventional displays. In this dissertation several challenges have been addressed from a perception-based point of view.

First, the mismatch of spatial resolution of common film cameras and (mostly much lower) display resolution has been identified. The first contribution of this thesis (Chapter 4) has been a method that enables apparent display resolution enhancement (ADRE) for arbitrary footage. This approach determines a spatio-temporal video transformation from saliency and gaze data for a given video in such a way that a previous resolution enhancement algorithm yields optimal results. The resolution enhancement effect exploits temporal summation in the foveal region during smooth pursuit eye motion to reconstruct high spatial detail in the retina. As a result, a commodity high frame-rate display is able to show arbitrary videos at higher perceived quality.

Second, the mismatch between camera blur and motion blur perceived in reality by the HVS motivated the contributions in Chapter 5. Gaze-aware perceptual blur model takes the estimated scan path of a given video into account and enables to recreate the amount of blur perceived naturally while watching the video. This way, temporal video artifacts such as judder and ghosting, perceived in the periphery when watching short exposure HFR videos, are removed. In addition, the approach can be used to subtly direct the user's gaze due to the property of the HVS to unconsciously follow regions of high detail. Another application is synthetic and virtual shutter simulation which can be pixel-precisely adjusted at interactive rates.

Over the last years, tremendous progress in head-mounted displays (HMD) has made virtual reality available for the consumer market. Immersion is the ultimate goal of VR headsets in order to produce a convincing user experience. However, the personal user experience is limited by insufficient or imprecise calibration functionality with current HMDs which often induce motion sickness. Active gaze tracking is not available in popular VR headsets. This prevents usage of gaze-aware applications in HMDs. Furthermore, it is difficult to learn about user behavior in virtual worlds. Although proprietary extensions are provided by a small number of companies, those components are prohibitively expensive and usability is limited due to closed-source implementations.

In the third part of this thesis, an affordable hardware and software solution for drift-free eye-tracking and user-friendly HMD calibration has been presented in Chapter 6. The novel modular binocular eye-tracking head-mounted display (ETHMD) relies on video-oculography for low-latency tracking of both eyes. The prototype supports personalizable lens positioning to accommodate for different interocular distances. The integrated mirrors split infrared light for eye tracking and, as a result, overcome the problem of a non-optimal viewing angle of the eye-tracking camera of previous

HMD hardware designs. The lean design of the presented prototype provides full FOV while using commodity cameras for eye tracking.

On the software side, a model-based calibration procedure adjusts the eye tracking system and gaze estimation to varying lens positions. Challenges such as partial occlusions due to the lens holders and eye lids are handled by a novel robust monocular pupil-tracking approach. The HMD design and the introduced algorithms constitute a low-latency VR system which is affordable and simple to calibrate. It is suited to the needs of immersive VR and gaze-contingent real-time display algorithms.

Measuring relative gaze direction opens the door to a much wider spectrum of gaze-aware VR applications and games when using HMDs. A selection of applications has been demonstrated in Chapter 6, such as gaze calibration, accommodation simulation, gaze control of virtual avatars, gaze map estimation and gaze analysis for immersive videos. Due to the fact that VR applications require high display refresh rates and a wide FOV to avoid motion sickness, rendering for the HMD is especially GPU-demanding.

For this reason, the fourth and last contribution in this dissertation focuses on gaze-contingent real-time rendering, presented in Chapter 7. Exploiting the decrease in perceivable display quality in the periphery is especially beneficial for wide field-of-view VR headsets. The novel perceptual sampling scheme adaptively adjusts shading quality to different limitations of the HVS, such as spatial acuity, brightness adaptation and temporal sensitivity. The approach is suited for modern physically-based shading models and fits into current rendering pipelines. In contrast to related approaches, the method is compatible with common GPU hardware and shading languages, making it well-suited for many platforms.

Current limitations of passive gaze-aware display algorithms arise from imperfect saliency prediction, e.g., for video content that features no or multiple salient regions. The precision of attention models being able to simulate and predict *selective attention* determines the success of gaze-aware video rendering methods. Our perceived environment is not only the sum of colors, angles, and motion. People and objects around us resonate with emotional meaning. For robust gaze prediction it is mandatory to understand how attention and empathy interact. Selective attention is determined both by goal-directed, top-down control as well as by bottom-up processing of all input channels. Unified perceptual models should, therefore, include more than just vision. Auditory and haptic perception also have a large impact when selecting attention. In addition, more research is needed into how we perceive with our peripheral vision and how to suitably render and display image information in the periphery. As displays become more sophisticated and advanced, a deeper understanding of our perception will be needed, including difficult-to-measure aspects such as visual comfort.

The fact that gaze-contingent methods concentrate largely on the single-viewer case constitutes another limitation. To accommodate several people looking at the same screen the presented methods need to be suitably adapted and extended.

8.2 Future Directions

Authentic visual realism arguably constitutes the strongest cue for our sensation of reality. By enabling immersive visual realism, the presented and future research will open up exciting new application scenarios for the use of gaze-contingent displays, not only in visual entertainment and gaming but also in areas like visualization and fundamental perception research.

The software-based approach for apparent display resolution enhancement has been only tested for videos so far. However, current VR headsets are mostly limited in terms of spatial display resolution. Hence, increasing perceived resolution can render highly beneficial in this area. Currently, the image optimization is computationally costly if applied as described in Chapter 4. However, as we perceive high frequency detail only in ≈1% of our visual field, the amount of pixels required to be processed is limited. A fast GPU implementation as suggested by Templin et al. could therefore be an option to exploit ADRE in real-time applications to overcome the currently limited spatial resolution in VR headsets [TDR+11].

Another important path of future work are ways for rendering perceptually convincing, viewpoint-correct synthetic worlds as well as captured videos at very high frame rates. In theory, AMOLED displays, which are primarily used for novel HMDs, enable refresh rates of more than 1000Hz, much higher than traditional video and application frame rates. High temporal display resolution brings perceived visual impression closer to what we perceive in reality. However, rendering for a wide FOV and at high frame rates is prohibitively expensive for complex scenes. Temporal reprojection of videos, such as presented for ADRE and the perceptual blur, or time warp features for real-time applications currently approximate in-between frames to increase the frame rate at much lower cost. However, effects on visual perception and motion sickness have not exhaustively been analyzed yet when presenting interpolated frames instead of viewpoint-correct frames. Especially wide FOV displays may benefit from gaze-aware temporal upscaling methods due to the strong decrease in visual performance towards the periphery.

Orthogonally, adjusting rendering quality to perceptual limits has proven to be beneficial, as shown in the perceptual sampling project. Although not tested for mobile VR headsets, such as the GearVR™, once mobile gaze tracking is available the proposed approach is directly applicable. In comparison to a desktop environment, the gaze-aware render method would probably result in an even higher performance boost on mobile hardware due to the fact that mobile GPU architectures are more sensitive to expensive shading calls. In addition, extended approaches could combine shading-based with geometry-based foveation schemes to maximize the decrease in work load.

The selection of gaze-aware applications presented in this dissertation, and the extent to which these methods enhance user experience, only form a small subset of what is possible if models of visual perception are included into the render pipeline. For example, the perceptual blur described in Chapter 5 assumes a reasonable but constant display brightness for estimating the amount of blur required for a natural video impression without artifacts. However, in reality spatio-temporal sensitivity depends on a number of factors such as environment luminance as well as local display

brightness which changes over time. Additionally, sensitivity depends on eccentricity of the stimulus and varies accordingly across the visual field. Although, the formulated perceptual blur model supports pixel-precise filtering, the mentioned properties of human vision with respect to *spatio-temporal* sensitivity (Chapter 2.3) have not been considered and remain to be explored in future research. An extended spatio-temporal model of the perceptual blur could take per-pixel display brightness, eccentricity information and eye adaptation into account to control the simulated summation duration as well as the amount of spatial blur.

Another wide field for future research is high-fidelity video rendering. The most striking limitations of current rendering approaches for immersive video are available bandwidth and the view location constraint. Gaze-aware video rendering and foveated video compression schemes are promising directions to achieve the bandwidth reduction necessary to raise video resolution to the acuity limit in the foveal and peripheral viewing areas. The practicability of streamable foveated video systems is currently limited by system latency [RYS+16]. This lag potentially results in a noticeable switch of streamed video resolution when the user rotates his head or performs a saccade. Therefore, novel methods for robustly predicting gaze direction several frame ahead are required to decrease the latency to an acceptable level. The second challenge is to overcome the limitation of not being able to change the viewpoint while watching a panoramic video. Being restricted to a static viewpoint results from the fact that immersive videos are currently captured by a single monocular panoramic camera. However, assuming a fixed head position limits achievable immersion. Desirably, a deeper degree of immersion could be achieved by facilitating ego-motion parallax as well as viewpoint-correct stereopsis during immersive video playback. Inspired by previous work on free-viewpoint video rendering [LKM14], one way could be to estimate depth and image correspondences from omni-directional, multi-stereo footage and apply wide-baseline reconstruction. However, to achieve this goal in real-time for VR headsets, a number of interdisciplinary challenges from video processing, computer graphics, and applied visual perception need to be addressed collectively.

We have only just begun to explore the possibilities of gaze-contingent computational displays. Many more exciting methods and applications are certain to be discovered in the coming years. With affordable eye tracking solutions becoming consumer electronics items, the widespread deployment of gaze-contingent VR displays is only a question of time and social acceptance.

References

[AB91] Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991.

[Abr13] Michael Abrash. Why virtual isn't real to your brain: judder, June 2013. http://blogs.valvesoftware.com/abrash/why-virtual-isnt-real-to-your-brain-judder/, vis. 09-26-2016.

[Abr14] Michael Abrash. What VR could, should, and almost certainly will be within two years. Presentation at Steam Dev Days 2014, February 2014. http://media.steampowered.com/apps/steamdevdays/slides/vrshouldbe.pdf, vis. 09-12-2016.

[Ade82] Edward H. Adelson. Saturation and adaptation in the rod system. *Vision research*, 22(10):1299–1312, 1982.

[AF57] Hermann Aubert and Richard Foerster. Untersuchungen über den Raumsinn der Retina. *Albrecht Von Graefe's Arch Klin Experiment Ophthalmol*, 3:1–37, 1857.

[AFB+13] Oleg Alexander, Graham Fyffe, Jay Busch, Xueming Yu, Ryosuke Ichikari, Andrew Jones, Paul Debevec, Jorge Jimenez, Etienne Danvoye, Bernardo Antionazzi, et al. Digital ira: Creating a real-time photoreal digital actor. In *ACM SIGGRAPH 2013 Posters*, page 1, 2013.

[AGI14] Agisoft PhotoScan, 2014. http://www.agisoft.com/, vis. 12-20-2014.

[AHEF02] Markus Andiel, Siegbert Hentschke, Thorsten Elle, and Eduard Fuchs. Eye tracking for autostereoscopic displays using web cams. In *Electronic Imaging 2002*, pages 200–206. International Society for Optics and Photonics, 2002.

[AHSS04] Aseem Agarwala, Aaron Hertzmann, David H. Salesin, and Steven M. Seitz. Keyframe-based tracking for rotoscoping and animation. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, pages 584–591, New York, NY, USA, 2004.

[AKLA11] Francis Heed Adler, Paul L. Kaufman, Leonard A. Levin, and Albert Alm. *Adler's Physiology of the Eye*. Elsevier Health Sciences, 2011.

[ARD15] Arduino microprocessor, 2015. https://www.arduino.cc/, vis. 09-12-2016.

[AU05] W. Allen and R. Ulichney. Wobulation: Doubling the addressed resolution of projection displays. In *Proc. of the International Symposium Digest of Technical Papers (SID)*, pages 1514–1517. Society for Information Display, 2005.

[BAHLC09] Alexandre Benoit, David Alleysson, Jeanny Herault, and Patrick Le Callet. *Spatio-temporal Tone Mapping Operator Based on a Retina Model*, pages 12–22. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[Bak49] Howard Dehaven Baker. The course of foveal light adaptation measured by the threshold intensity increment. *JOSA*, 39(2):172–179, 1949.

[Bas06] Adolf Basler. Über das Sehen von Bewegungen. *Pflügers Archiv European Journal of Physiology*, 115(11):582–601, 1906.

[BB09] S. Basu and P. Baudisch. System and process for increasing the apparent resolution of a display, 2009. US Patent 7,548,662.

[BBK+15] Amit Bermano, Thabo Beeler, Yeara Kozlov, Derek Bradley, Bernd Bickel, and Markus Gross. Detailed spatio-temporal reconstruction of eyelids. *ACM Transactions on Graphics (TOG)*, 34(4):44, 2015.

[BBS01] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *CVPR 2001. Proceedings*, volume 1, pages I–355. IEEE, 2001.

[BCFW08] Dirk Bartz, Douglas Cunningham, Jan Fischer, and Christian Wallraven. The role of perception for computer graphics. *Eurographics state-of-the-art-reports*, pages 65–86, 2008.

[BCKD15] Kenan Bektas, Arzu Cöltekin, Jens Krüger, and Andrew T. Duchowski. A testbed combining visual perception models for geographic gaze contingent displays. In *Eurographics Conference on Visualization (EuroVis)-Short Papers*, 2015.

[BDB+06] Erhardt Barth, Michael Dorr, Martin Böhme, Karl Gegenfurtner, and Thomas Martinetz. Guiding the mind's eye: improving communication and vision by external control of the scanpath. In *Electronic Imaging 2006*, pages 60570D–60570D. International Society for Optics and Photonics, 2006.

[BDK+06] M. Boehme, M. Dorr, C. Krause, T. Martinetz, and E. Barth. Eye movement predictions on natural videos. *Neurocomputing*, pages 1996–2004, 2006.

[BDMB06] Martin Böhme, Michael Dorr, Thomas Martinetz, and Erhardt Barth. Gaze-contingent temporal filtering of video. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, pages 109–115. ACM, 2006.

[BEM11] Pablo Bauszat, Martin Eisemann, and Marcus Magnor. Guided image filtering for interactive high-quality global illumination. *Computer Graphics Forum*, 30(4):1361–1368, 2011.

[BF12a] Floraine Berthouzoz and Raanan Fattal. Apparent resolution enhancement for motion videos. In *Proceedings of ACM Symposium on Applied Perception (SAP)*, pages 91–98, 2012.

[BF12b] Floraine Berthouzoz and Raanan Fattal. Resolution enhancement by vibrating displays. *ACM Transactions on Graphics (TOG)*, pages 15:1–15:14, 2012.

[BHSS15] Chrisitan Bailer, José Henriques, Norbert Schmitz, and Didier Stricker. A simple real-time eye tracking and calibration approach for autostereoscopic 3d displays. *Proceedings of 10th International Conference on Computer Vision, Theory and Applications 2015*, March 2015.

[BI13] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):185–207, 2013.

[BJ11] Peter Bazanov and Toni Järvenpää. Gaze estimation for near-eye display based on fusion of starburst algorithm and fern natural features. In *FRUCT 2011*, pages 1–8, 2011.

[BJD+15] Zoya Bylinskii, Tilke Judd, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT saliency benchmark, March 2015. http://saliency.mit.edu/, vis. 12-09-2016.

[BK08] Brian A. Barsky and Todd J. Kosloff. Algorithms for rendering depth of field effects in computer graphics. In *Proceedings of the 12th WSEAS International Conference on Computers*, ICCOMP'08, pages 999–1010a, Stevens Point, Wisconsin, USA, 2008.

[BKBM04] Martin Böhme, Christopher Krause, Erhardt Barth, and Thomas Martinetz. Eye movement predictions enhanced by saccade detection. In *In: Brain Inspired Cognitive Systems*, pages 1–7, 2004.

[BKLJP04] Doug A. Bowman, Ernst Kruijff, Joseph J. LaViola Jr., and Ivan Poupyrev. *3DUI: theory and practice*. Addison-Wesley, 2004.

[BKM05] Josephine Battista, Michael Kalloniatis, and Andrew Metha. Visual function: the problem with eccentricity. *Clinical and Experimental Optometry*, 88(5):313–321, 2005.

[BKR+14] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl. State-of-the-art of visualization for eye tracking data. In *Proceedings of EuroVis*, 2014.

[Blo85] Adolphe-Moïse Bloch. Experiences sur la vision. *CR Seances Soc. Biol. Paris*, 37:493–495, 1885.

[BM97] D. C. Burr and M. J. Morgan. Motion deblurring in human vision. *Proc. Biol. Sci.*, 264(1380):431–436, 1997.

[BMS02] G. Boccignone, A. Marcelli, and G. Somma. Analysis of dynamic scenes based on visual attention. *Proc. of IEEE Workshop on Artificial Intelligence for Industrial Applications (AIIA)*, pages 1–10, 2002.

[BMSG09] Reynold Bailey, Ann McNamara, Nisha Sudarsanam, and Cindy Grimm. Subtle gaze direction. *ACM Transactions on Graphics (TOG)*, 28(4):100, 2009.

[Bou10] Jean-Yves Bouguet. Camera Calibration Toolbox for Matlab, 2010. http://www.vision.caltech.edu/bouguetj/calib_doc/, vis. 09-12-2016.

[BP94] Shumeet Baluja and Dean Pomerleau. Non-intrusive gaze tracking using artificial neural networks. Technical report, Carnegie Mellon University, Pittsburgh, PA, USA, 1994.

[Bro02] B. Brown. *Cinematography: Theory and Practice : Imagemaking for Cinematographers, Directors & Videographers*. Focal Press, 2002.

[BS14] Jennifer Romano Bergstrom and Andrew Schall. *Eye Tracking in User Experience Design*. Morgan Kaufmann Publishers Inc., 2014.

[BSCB00] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 417–424, 2000.

[Bur81] D. C. Burr. Temporal summation of moving images by the human visual system. volume 211, pages 321–339. The Royal Society, 1981.

[CA84] Nancy J. Coletta and Anthony J. Adams. Rod-cone interaction in flicker detection. *Vision Research*, 24(10):1333–1340, 1984.

[CBN05] Hannah Faye Chua, Julie E Boland, and Richard E Nisbett. Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12629–12633, 2005.

[CCW03] Kirsten Cater, Alan Chalmers, and Greg Ward. Detail to attention: exploiting visual tasks for selective rendering. In *ACM International Conference Proceeding Series*, volume 44, pages 270–280, 2003.

[CDdS06] Alan Chalmers, Kurt Debattista, and Luis Paulo dos Santos. Selective rendering: Computing only what you see. In *Proceedings of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*, GRAPHITE '06, pages 9–18, 2006.

[CDF+06] Forrester Cole, Doug DeCarlo, Adam Finkelstein, Kenrick Kin, Keith Morley, and Anthony Santella. Directing gaze in 3D models with stylized focus. *Eurographics Symposium on Rendering*, pages 377–387, June 2006.

[CE14] Siyuan Chen and J. Epps. Efficient and robust pupil size and blink estimation from near-field video sequences for human machine interaction. *IEEE Transactions on Cybernetics*, 44(12):2356–2367, Dec 2014.

[Che02] Milton Chen. Leveraging the asymmetric sensitivity of eye contact for videoconference. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 49–56. ACM, 2002.

[CHEK08] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. *Advances in neural information processing systems*, 2008.

[CHH+09] M. Cerf, J. Harel, A. Huth, W. Einhäuser, and C. Koch. Decoding what people see from where they look: Predicting visual stimuli from scanpaths. *Attention in Cognitive Systems*, pages 15–26, 2009.

[CLD14] Dario Cazzato, Marco Leo, and Cosimo Distante. An investigation on the feasibility of uncalibrated and unconstrained gaze tracking for human assistive applications by using head pose estimation. *Sensors*, 14(5):8363–8379, 2014.

[CLR04] Marisa Carrasco, Sam Ling, and Sarah Read. Attention alters appearance. *Nature neuroscience*, 7(3):308–313, 2004.

[CPC84] Robert L. Cook, Thomas Porter, and Loren Carpenter. Distributed ray tracing. *Computer Graphics*, 18(3):137, 1984.

[CR74] A. Cowey and E. T. Rolls. Human cortical magnification factor and its relation to visual acuity. *Experimental Brain Research*, 21(5):447–454, 1974.

[CS02] Maurizio Corbetta and Gordon L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002.

[CSKH90] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson. Human photoreceptor topography. *Journal of Comparative Neurology*, pages 497–523, 1990.

[CT82] Robert L. Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (TOG)*, 1(1):7–24, 1982.

[CTCS00] Jin-Xiang Chai, Xin Tong, Shing-Chow Chan, and Heung-Yeung Shum. Plenoptic sampling. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 307–318, 2000.

[CTM13] Petrik Clarberg, Robert Toth, and Jacob Munkberg. A sort-based deferred shading architecture for decoupled sampling. *ACM Transactions on Graphics (TOG)*, 32(4):141, 2013.

[CW11] Douglas W. Cunningham and Christian Wallraven. *Experimental design: From user studies to psychophysics*. CRC Press, 2011.

[CWLZ13] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013.

[CY13] H. R. Chennamma and Xiaohui Yuan. A survey on eye-gaze tracking techniques. *Indian Journal of Computer Science & Engineering (IJCSE)*, 4(5):388, October 2013.

[Dal98] Scott J Daly. Engineering observations from spatiovelocity and spatiotemporal visual models. In *Photonics West'98 Electronic Imaging*, pages 180–191. International Society for Optics and Photonics, 1998.

[DBBS06] Thomas Dera, Guido Boning, Stanislavs Bardins, and Erich Schneider. Low-latency video tracking of horizontal, vertical, and torsional eye movements as a basis for 3dof real-time motion control of a head-mounted camera. In *SMC'06, Proceedings*, volume 6, pages 5191–5196. IEEE, 2006.

[DBMB06] Michael Dorr, Martin Böhme, Thomas Martinetz, and Erhardt Barth. Gaze-contingent spatio-temporal filtering in a head-mounted display. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 205–207. Springer, 2006.

[DBS+09] Andrew T. Duchowski, David Bate, Paris Stringfellow, Kaveri Thakur, Brian J. Melloy, and Anand K. Gramopadhye. On spatiochromatic visual sensitivity and peripheral color lod management. *ACM Transactions on Applied Perception (TAP)*, 6(2):9, 2009.

[DÇ07] Andrew T. Duchowski and Arzu Çöltekin. Foveated gaze-contingent displays for peripheral lod management, 3d visualization, and stereo imaging. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(4):6, 2007.

[DCK13] Michal Dziemianko, Alasdair Clarke, and Frank Keller. Object-based saliency as a predictor of attention in visual tasks. In *Proc. of the 35th Conference of the Cognitive Science Society*, pages 2237–2242, 2013.

[DCM04] Andrew T. Duchowski, Nathan Cournia, and Hunter Murphy. Gaze-contingent displays: A review. *CyberPsychology & Behavior*, 7(6):621–634, 2004.

[Dem04] Joe Demers. Depth of field: A survey of techniques. *GPU Gems*, 1(375):390–400, 2004.

[DER+10a] P. Didyk, E. Eisemann, T. Ritschel, K. Myszkowski, and H.P. Seidel. Apparent display resolution enhancement for moving images. *ACM Transactions on Graphics (TOG)*, pages 113:1–113:8, 2010.

[DER+10b] Piotr Didyk, Elmar Eisemann, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel. Perceptually-motivated real-time temporal upsampling of 3d content for high-refresh-rate displays. In *Computer Graphics Forum*, volume 29, pages 713–722, 2010.

[DFRR10] Jennifer E. Doble, Debby L. Feinberg, Mark S. Rosner, and Arthur J. Rosner. Identification of binocular vision dysfunction in traumatic brain injury patients and effects of individualized prismatic spectacle lenses. *PM&R*, 2(4):244–253, 2010.

[DGY07] Andreas Dietrich, Enrico Gobbetti, and Sung-Eui Yoon. Massive-model rendering techniques. *IEEE Computer Graphics and Applications*, 27(6):20–34, 2007.

[DHG+14] Andrew T. Duchowski, Donald H. House, Jordan Gestring, Rui I. Wang, Krzysztof Krejtz, Izabela Krejtz, Radosław Mantiuk, and Bartosz Bazyluk. Reducing visual discomfort of 3d stereoscopic displays with gaze-contingent depth-of-field. In *Proceedings of the ACM Symposium on Applied Perception (SAP)*, pages 39–46, 2014.

[DI03] Nitin Dhavale and Lativent Itti. Saliency-based multifoveated mpeg compression. In *Signal processing and its applications, 2003. Proceedings. Seventh international symposium on*, volume 1, pages 229–232. IEEE, 2003.

[DMGB10] Michael Dorr, Thomas Martinetz, Karl R. Gegenfurtner, and Erhardt Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, 10, 2010.

[DS02a] Doug DeCarlo and Anthony Santella. Stylization and abstraction of photographs. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 769–776, 2002.

[DS02b] Doug DeCarlo and Anthony Santella. Stylization and abstraction of photographs. *ACM Transactions on Graphics (TOG)*, 21(3):769–776, 2002.

[DSR+00] Andrew T. Duchowski, Vinay Shivashankaraiah, Tim Rawls, Anand K. Gramopadhye, Brian J. Melloy, and Barbara Kanki. Binocular eye tracking in virtual reality for inspection training. In *ETRA'00, Proceedings*, pages 89–96. ACM, 2000.

[Duc79] Claude E. Duchon. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology*, 18(8):1016–1022, 1979.

[Duc02] Andrew T. Duchowski. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, 34(4):455–470, 2002.

[Duc07] Andrew T. Duchowski. *Eye tracking methodology: Theory and practice*, volume 373. Springer, 2007.

[DVB12] Michael Dorr, Eleonora Vig, and Erhardt Barth. Eye movement prediction and variability on natural video data sets. *Visual cognition*, 20(4-5):495–514, 2012.

[DVC09] N. Damera-Venkata and N.L. Chang. Display supersampling. *ACM Transactions on Graphics (TOG)*, pages 9:1–9:19, 2009.

[DVDV93] Russell L. De Valois and Karen K. De Valois. A multi-stage color model. *Vision research*, 33(8):1053–1065, 1993.

[DW61] P. M. Daniel and D. Whitteridge. The representation of the visual field on the cerebral cortex in monkeys. *The Journal of physiology*, 159(2):203–221, 1961.

[DWB06] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *Robotics & Automation Magazine, IEEE*, 13(2):99–110, 2006.

[EE16] Operating Eurovision and Euroradio (EBU). Safe areas for 16:9 television production. Technical report, 2016. https://tech.ebu.ch/docs/r/r095.pdf, vis. 09-16-2016.

[EJGAC+15] David E. Jacobs, Orazio Gallo, Emily A. Cooper, Kari Pulli, and Marc Levoy. Simulating the visual experience of very bright and very dark scenes. *ACM Transactions on Graphics (TOG)*, 34(3):25, 2015.

[ENSB13] Christian Eisenacher, Gregory Nichols, Andrew Selle, and Brent Burley. Sorted deferred shading for production path tracing. *Comput. Graph. Forum*, 32(4):125–132, 2013.

[EPI15] Epic Games, Inc., Unreal Engine, 2015. http://www.unrealengine.com/, vis. 12-09-2016.

[ERK08] Wolfgang Einhäuser, Ueli Rutishauser, and Christof Koch. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, 8(2):2–2, 2008.

[EUWM13] Gabriel Eilertsen, Jonas Unger, Robert Wanat, and Rafał Mantiuk. Survey and evaluation of tone mapping operators for hdr video. In *ACM SIGGRAPH 2013 Talks*, page 11, 2013.

[Fai13] Mark D. Fairchild. *Color appearance models*. John Wiley & Sons, 2013.

[Fai15] Mark D. Fairchild. Seeing, adapting to, and reproducing the appearance of nature. *Applied optics*, 54(4):B107–B116, 2015.

[FBA+94] Henry Fuchs, Gary Bishop, Kevin Arthur, Leonard McMillan, Ruzena Bajcsy, Sang Lee, Hany Farid, and Takeo Kanade. Virtual space teleconferencing using a sea of cameras. In *Proc. First International Conference on Medical Robotics and Computer Assisted Surgery*, volume 26, 1994.

[FCW+10] Martin Fuchs, Tongbo Chen, Oliver Wang, Ramesh Raskar, Hans-Peter Seidel, and Hendrik P.A. Lensch. Real-time temporal shaping of high-speed video streams. *Computers & Graphics*, 34(5):575–584, 2010.

[Fen06] Xiao-Fan Feng. Cd motion-blur analysis, perception, and reduction using synchronized backlight flashing. In *Electronic Imaging 2006*, pages 60570M–60570M. International Society for Optics and Photonics, 2006.

[Fen14] Wesley Fenlon. 48 FPS and Beyond: How High Frame Rate Films Affect Perception, 2014. http://www.tested.com/art/movies/452387-48-fps-and-beyond-how-high-frame-rates-affect-perception/, vis. 09-26-2016.

[FF+96] Andrew W. Fitzgibbon, Robert B. Fisher, et al. A buyer's guide to conic fitting. *DAI Research paper*, 1996.

[FFIKP07] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7(1):10–10, 2007.

[FH14] Masahiro Fujita and Takahiro Harada. Foveated real-time ray tracing for virtual reality headset. Technical report, Light Transport Entertainment Inc., 2014.

[FMR08] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[FP04] Jean-Philippe Farrugia and Bernard Péroche. A progressive rendering algorithm using an adaptive perceptually based image metric. In *Computer Graphics Forum*, volume 23, pages 605–614, 2004.

[FPSG96] James A. Ferwerda, Sumanta N. Pattanaik, Peter Shirley, and Donald P. Greenberg. A model of visual adaptation for realistic image synthesis. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pages 249–258. ACM, 1996.

[FR84] Burkhart Fischer and E. Ramsperger. Human express saccades: extremely short reaction times of goal directed eye movements. *Experimental Brain Research*, 57(1):191–195, 1984.

[FR99] Elisabeth M. Fine and Gary S. Rubin. Reading with simulated scotomas: attending to the right is better than attending to the left. *Vision Research*, 39(5):1039–1048, 1999.

[FWMG15] Simone Frintrop, Thomas Werner, and German Martin Garcia. Traditional saliency reloaded: a good old model in new shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 82–90, 2015.

[Gam16] Epic Games. Introducing real-time cinematography: Hellblade. SIGGRAPH Best Real-time Graphics Award, 2016. https://www.youtube.com/watch?v=KeNXEjNkEs0, vis. 09-12-2016.

[GDS14] Steven Galea, Kurt Debattista, and Sandro Spina. Gpu-based selective sparse sampling for interactive high-fidelity rendering. In *Games and Virtual Worlds for Serious Applications (VS-GAMES), 2014 6th International Conference on*, pages 1–8. IEEE, 2014.

[GFD+12] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. Foveated 3d graphics. *ACM Transactions on Graphics (TOG)*, 31(6):164, 2012.

[GGSC96] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pages 43–54. ACM, 1996.

[GH30] Ragnar Granit and Phyllis Harper. Comparative studies on the peripheral and central retina. *American Journal of Physiology–Legacy Content*, 95(1):211–228, 1930.

[Gla99] Andrew Glassner. An open and shut case. *IEEE Comput. Graph. Appl.*, 19(3):82–92, May 1999.

[Gol09] E. Bruce Goldstein. *Encyclopedia of perception*. Sage Publications, 2009.

[Gol13] E. Bruce Goldstein. *Sensation and perception*. Cengage Learning, 2013.

[GOSG97] S. J. Galvin, R. P. O'Shea, A. M. Squire, and D. G. Govan. Sharpness overconstancy in peripheral vision. *Vision Res*, 37(15):2035–2044, 1997.

[GP98] W.S. Geisler and J.S. Perry. A real-time foveated multi-resolution system for low-bandwidth video communication. In *Human Vision and Electronic Imaging, SPIE Proceedings*, number 3299, pages 294–305, 1998.

[GP99] W.S. Geisler and J.S. Perry. Variable-resolution displays for visual communication and simulation. In *The Society for Information Display*, number 30, pages 420–423, 1999.

[GPN06] Wilson S. Geisler, Jeffrey S. Perry, and Jiri Najemnik. Visual search: The role of peripheral information measured using gaze-contingent displays. *Journal of Vision*, 6(9):1–1, 2006.

[Gre70] Daniel G. Green. Regional variations in the visual acuity for interference fringes on the retina. *The Journal of physiology*, 207(2):351–356, 1970.

[GSV+03] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M Angela Sasse. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 529–536. ACM, 2003.

[GVC03] R. Gaborski, Vishal S. Vaingankar, and R. L. Canosa. Goal directed visual search based on color cues: Cooperative effects of top-down & bottom-up visual attention. *Proceedings of the Artificial Neural Networks in Engineering, Rolla, Missouri*, 13:613–618, 2003.

[Har16] Carlo Harvey. *Multi-Modal Perception for Selective Rendering*. Phd, University of Warwick, November 2016.

[HBCM07] John M. Henderson, James R. Brockmole, Monica S. Castelhano, and Michael Mack. Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye movements: A window on mind and brain*, pages 537–562, 2007.

[HCOB10] Robert T. Held, Emily A. Cooper, James F. O'brien, and Martin S. Banks. Using blur to affect perceived distance and size. *ACM Transactions on Graphics (TOG)*, 29(2), 2010.

[HCS10] Jasminka Hasic, Alan Chalmers, and Elena Sikudova. Perceptually guided high-fidelity rendering exploiting movement bias in visual attention. *ACM Transactions on Applied Perception (TAP)*, 8(1):6:1–6:19, 2010.

[HEKR14] Mike Horsley, Matt Eliot, Bruce Allen Knight, and Ronan Reilly. *Current trends in eye tracking research*. Springer, 2014.

[HF86] Donald C. Hood and Marcia A. Finkelstein. Sensitivity to light. *Handbook of Perception and Human Performance (Vol. 1: Sensory Processes and Perception). John Wiley and Sons, New York.*, 1986.

[HGAB08] David M. Hoffman, Ahna R. Girshick, Kurt Akeley, and Martin S. Banks. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision*, 8(3):33–33, 2008.

[HGF14] Yong He, Yan Gu, and Kayvon Fatahalian. Extending the graphics pipeline with adaptive, multi-rate shading. *ACM Transactions on Graphics (TOG)*, 33(4):Article–142, 2014.

[HGG98] Stephen T. Hammett, Mark A. Georgeson, and Andrei Gorea. Motion blur and motion sharpening: temporal smear and local contrast non-linearity. *Vision research*, 38(14):2099–2108, 1998.

[HHQ+13] Junwei Han, Sheng He, Xiaoliang Qian, Dongyang Wang, Lei Guo, and Tianming Liu. An object-oriented visual saliency detection framework based on sparse coding representations. *IEEE transactions on circuits and systems for video technology*, 23(12):2009–2021, 2013.

[HJ10] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(3):478–500, 2010.

[HKB11] David M. Hoffman, Vasiliy I. Karasev, and Martin S. Banks. Temporal presentation protocols in stereoscopic displays: Flicker visibility, perceived motion, and perceived depth. *Journal of the Society for Information Display*, 19(3):271, 2011.

[HKP07] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, pages 545–552, 2007.

[HLCC08] Sébastien Hillaire, Anatole Lécuyer, Rémi Cozot, and Géry Casiez. Using an eye-tracking system to improve camera motions and depth-of-field blur effects in virtual environments. In *IEEE Virtual Reality Conference (VR '08)*, pages 47–50, 2008.

[HLSR14] John M. Henderson, Steven G. Luke, Joseph Schmidt, and John E. Richards. Co-registration of eye movements and event-related potentials in connected-text paragraph reading. *Eye movement-related brain activity during perceptual and cognitive processing*, page 67, 2014.

[HLW15] Fu-Chung Huang, David Luebke, and Gordon Wetzstein. The light field stereoscope. *ACM SIGGRAPH Emerging Technologies*, 24, 2015.

[HM39] Selig Hecht and Esther U. Mintz. The visibility of single lines at various illuminations and the retinal basis of visual resolution. *The Journal of general physiology*, 22(5):593–612, 1939.

[HMY13] Takahiro Harada, Jay McKee, and Jason C. Yang. Forward+: A step toward film-style shading in real time. *GPU Pro*, 4:115–134, 2013.

[HMYS01] Jörg Haber, Karol Myszkowski, Hitoshi Yamauchi, and Hans-Peter Seidel. Perceptually guided corrective splatting. In *Computer Graphics Forum*, volume 20, pages 142–153, 2001.

[HNA+11] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[HNM12] Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. Eye tracker data quality: what it is and how to measure it. In *ETRA'12, Proceedings*, pages 45–52. ACM, 2012.

[Hop98] Hugues Hoppe. Smooth view-dependent level-of-detail control and its application to terrain rendering. In *Visualization'98. Proceedings*, pages 35–42. IEEE, 1998.

[HS81] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[HS00] Z. Hara and N Shiramatsu. Improvement in the picture quality of moving pictures for matrix displays. *Journal of the Society for Information Display*, pages 129–137, 2000.

[HSH10] Liang Hu, Pedro V Sander, and Hugues Hoppe. Parallel view-dependent level-of-detail control. *Visualization and Computer Graphics, IEEE Transactions on*, 16(5):718–728, 2010.

[HVDFF14] John F. Hughes, Andries Van Dam, James D. Foley, and Steven K. Feiner. *Computer graphics: principles and practice*. Pearson Education, 2014.

[HZ03] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[IK01] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[IKN98] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.

[Jac91] Robert J. K. Jacob. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems (TOIS)*, 9(2):152–169, 1991.

[JDT12] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. *MIT Computer Science and Artificial Intelligency Laboratory Technical Report*, 2012. MIT-CSAIL-TR-2012-001.

[JEDT09] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Proc. of IEEE Conference on Computer Vision (ICCV)*, pages 2106–2113. IEEE, 2009.

[JFY+11] Andrew Jones, Graham Fyffe, Xueming Yu, Wan-Chun Ma, Jay Busch, Ryosuke Ichikari, Mark Bolas, and Paul Debevec. Head-mounted photometric stereo for performance capture. In *Visual Media Production (CVMP), 2011 Conference for*, pages 158–164. IEEE, 2011.

[JOK09] Lina Jansen, Selim Onat, and Peter König. Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1):29–29, 2009.

[JSIS+08] J Adam Jones, J Edward Swan II, Gurjot Singh, Eric Kolstad, and Stephen R Ellis. The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization (SAP)*, pages 9–14. ACM, 2008.

[JTST10] Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. MovieReshape: tracking and reshaping of humans in videos. *ACM Transactions on Graphics (TOG)*, pages 148:1–148:10, 2010.

[KAB15] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *arXiv preprint arXiv:1510.02927*, 2015.

[KBB+08] Stefan Kohlbecher, Stanislavs Bardinst, Klaus Bartl, Erich Schneider, Tony Poitschke, and Markus Ablassmeier. Calibration-free eye tracking by reconstruction of the pupil ellipse in 3d space. In *ETRA'08, Proceedings*, pages 135–138. ACM, 2008.

[KDM+16] Petr Kellnhofer, Piotr Didyk, Karol Myszkowski, Mohamed M Hefeeda, Hans-Peter Seidel, and Wojciech Matusik. Gazestereo3d: seamless disparity manipulations. *ACM Transactions on Graphics (TOG)*, 35(4):68, 2016.

[Kel61] D. H. Kelly. Visual Responses to Time-Dependent Stimuli - Amplitude Sensitivity Measurements†. *JOSA*, 51(4):422–429, 1961.

[Kel79] D. H. Kelly. Motion and Vision - Stabilized Spatio-temporal Threshold Surface. *J. Opt. Soc. Am.*, 69(10):1340–1349, 1979.

[KFSW09] Wolf Kienzle, Matthias O. Franz, Bernhard Schölkopf, and Felix A. Wichmann. Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of vision*, 9(5):7–7, 2009.

[KG13] Brian Karis and Epic Games. Real Shading in Unreal Engine 4. *SIGGRAPH Course Notes*, 2013. Physically Based Shading in Theory and Practice.

[KG14] Brian Karis and Epic Games. High quality temporal supersampling. *SIGGRAPH Course Notes*, 2014. Physically Based Shading in Theory and Practice.

[KHKH10] Frank Klefenz, Peter Husar, Daniel Krenzer, and Albrecht Hess. Real-time calibration-free autonomous eye tracker. In *ICASSP'10, Proceedings*, pages 762–765. IEEE, 2010.

[KHN16] Pawel Kasprowski, Katarzyna Harezlak, and Michał Niezabitowski. Eye movement tracking as a new promising modality for human computer interaction. In *2016 17th International Carpathian Control Conference (ICCC)*, pages 314–318. IEEE, 2016.

[KKK+16] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2176–2184, 2016.

[KKS09] Nishant Kumar, Stefan Kohlbecher, and Erich Schneider. A novel approach to video-based pupil tracking. In *SMC'09, Proceedings*, pages 1255–1262. IEEE, 2009.

[KL07] M. Kalloniatis and C. Luu. Temporal resolution, 2007. http://webvision.med.utah.edu/temporal.html, vis. 09-12-2016.

[KMH01] Robert Kosara, Silvia Miksch, and Helwig Hauser. Semantic depth of field. In *Proc. of the IEEE Symposium on Information Visualization*, pages 97–104, 2001.

[KMH+02] Robert Kosara, Silvia Miksch, Helwig Hauser, Johann Schrammel, Verena Giller, and Manfred Tscheligi. Useful properties of semantic depth of field for better f+ c visualization. In *ACM International Conference Proceeding Series*, volume 22, pages 205–210, 2002.

[KMS05] Grzegorz Krawczyk, Karol Myszkowski, and Hans-Peter Seidel. Perceptual effects in real-time tone mapping. In *Proceedings of the 21st ACM Spring Conference on Computer Graphics*, pages 195–202, 2005.

[KNC16] Christoph Kubisch and Nvidia Corporation. NVidia: Life of triangle - NVIDIA's logical pipeline. Nvidia Website, 2016. https://goo.gl/kmHvp1, vis. 03-21-2016.

[KNFJ14] Helga Kolb, Ralph Nelson, Eduardo Fernandez, and Bryan William Jones. Webvision - the organization of the retina and visual system, 2014. http://webvision.med.utah.edu/, vis. 10-20-2014.

[Kra16] Gregory Kramida. Resolving the vergence-accommodation conflict in head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(7):1912–1931, 2016.

[KSR+03] Alan Kingstone, Daniel Smilek, Jelena Ristic, Chris Kelland Friesen, and John D. Eastwood. Attention, researchers! It is time to take a look at the real world. *Current Directions in Psychological Science*, 12(5):176–180, 2003.

[KTB14] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

[KWB14] Matthias Kümmerer, Thomas Wallis, and Matthias Bethge. How close are we to understanding image-based saliency? *arXiv preprint arXiv:1409.7686*, 2014.

[LDC06] Peter Longhurst, Kurt Debattista, and Alan Chalmers. A gpu based saliency map for high-fidelity selective rendering. In *Proceedings of the ACM 4th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, pages 21–29, 2006.

[Lee09] Seung-Hyun Lee. Natural stereo images from a glasses-free, 3d display. *Newsroom of International Society for Optics and Photonics (SPIE)*, 2009. 10.1117/2.1201111.003916.

[LH01] David Luebke and Benjamin Hallen. Perceptually driven simplification for interactive rendering. In *Proceedings of the 12th Eurographics conference on Rendering*, pages 223–234. Eurographics Association, 2001.

[LHC10] Sheng Liu, Hong Hua, and Dewen Cheng. A novel prototype for an optical see-through head-mounted display with addressable focus cues. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 16(3):381–393, 2010.

[LHH+09] Gordon Love, David Hoffman, Philip Hands, James Gao, Andrew Kirby, and Martin Banks. High-speed switchable lens enables the development of a volumetric stereoscopic display. *Optics express*, 17(18):15716–15725, 2009.

[LK80] Denis N Lee and H Kalmus. The optic flow field: The foundation of vision [and discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 290(1038):169–179, 1980.

[LKA85] Dennis M. Levi, Stanley A. Klein, and A. P. Aitsebaomo. Vernier acuity, crowding and cortical magnification. *Vision research*, 25(7):963–977, 1985.

[LKM14] Christian Lipski, Felix Klose, and Marcus Magnor. Correspondence and depth-image based rendering: a hybrid approach for free-viewpoint video. *IEEE Trans. Circuits and Systems for Video Technology (T-CSVT)*, 24(6):942–951, June 2014.

[LLN+10] Christian Lipski, Christian Linz, Thomas Neumann, Markus Wacker, and Marcus Magnor. High resolution image correspondences for video post-production. In *Proc. European Conference on Visual Media Production (CVMP) 2010*, volume 7, pages 33–39. IEEE Computer Society, 2010.

[LM91] Michael S. Landy and J. Anthony Movshon. *Computational models of visual processing*. MIT press, 1991.

[LM00] Lester C. Loschky and George W. McConkie. User performance with gaze contingent multiresolutional displays. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, pages 97–103, 2000.

[LPB98] Sanghoon Lee, Marios S. Pattichis, and Alan C. Bovik. Rate control for foveated mpeg/h. 263 video. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, volume 2, pages 365–369. IEEE, 1998.

[LPB01] Sanghoon Lee, Marios S. Pattichis, and Alan Conrad Bovik. Foveated video compression with optimal rate control. *Image Processing, IEEE Transactions on*, 10(7):977–992, 2001.

[LPB02] Sanghoon Lee, Marios S. Pattichis, and Alan C. Bovik. Foveated video quality assessment. *IEEE Transactions on Multimedia*, 4(1):129–132, 2002.

[LRP+06] Justin Lairda, Mitchell Rosen, Jeff Pelz, Ethan Montag, and Scott Daly. Spatio-Velocity CSF as a function of retinal velocity using unstabilized stimuli. In *Human Vision and Electronic Imaging XI*, pages 32–43. SPIE, 2006.

[LSC04] Patrick Ledda, Luis Paulo Santos, and Alan Chalmers. A local model of eye adaptation for high dynamic range images. In *Proceedings of the 3rd International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, pages 151–160. ACM, 2004.

[LSF+15] Thomas Löwe, Michael Stengel, Emmy-Charlotte Förster, Steve Grogorick, and Marcus Magnor. Visualization and analysis of head movement and gaze data for immersive video in head-mounted displays. In *Proc. Workshop on Eye Tracking and Visualization (ETVIS) 2015*, volume 1, 2015.

[LTO+15] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (TOG)*, 34(4):47, 2015.

[LU13] Robert Gabriel Lupu and Florina Ungureanu. A survey of eye tracking methods and applications. Technical report, Technical University of Iaşi, 2013.

[Lue03] David P. Luebke. *Level of detail for 3D graphics*. Morgan Kaufmann, 2003.

[LW90] Marc Levoy and Ross Whitaker. Gaze-directed volume rendering. *ACM SIGGRAPH Computer Graphics*, 24(2):217–223, 1990.

[LW07] Lester C. Loschky and Gary S. Wolverton. How late can you update gaze-contingent multiresolutional displays without detection? *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(4):7, 2007.

[LWP05] Dongheng Li, David Winfield, and Derrick J. Parkhurst. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *CVPR 2005, Proceedings of*, pages 79–79. IEEE, 2005.

[LZDY08] Tie Liu, Nanning Zheng, Wei Ding, and Zejian Yuan. Video attention: Learning to detect a salient object sequence. In *International Conference on Pattern Recognition*, pages 1–4, 2008.

[MAX15] Next Limit S.L., Maxwell Render, 2015. http://www.maxwellrender.com, vis. 03-12-2015.

[MBG08] Ann McNamara, Reynold Bailey, and Cindy Grimm. Improving search task performance using subtle gaze direction. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization*, APGV '08, pages 51–56, 2008.

[MBM13] Radoslaw Mantiuk, Bartosz Bazyluk, and Rafal K. Mantiuk. Gaze-driven object tracking for real time rendering. In *Computer Graphics Forum*, volume 32, pages 163–173, 2013.

[MBS+12] Ann McNamara, Thomas Booth, Srinivas Sridharan, Stephen Caffey, Cindy Grimm, and Reynold Bailey. Directing gaze in narrative art. In *Proceedings of the ACM Symposium on Applied Perception*, pages 63–70, 2012.

[MBT11] Rafal Mantiuk, Bartosz Bazyluk, and Anna Tomaszewska. Gaze-dependent depth-of-field effect rendering in virtual environments. In *International Conference on Serious Games Development and Applications*, pages 1–12. Springer, 2011.

[MCNV14] Michael Mauderer, Simone Conte, Miguel A Nacenta, and Dhanraj Vishwanath. Depth perception with gaze-contingent depth of field. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 217–226, 2014.

[MD01] Hunter Murphy and Andrew T. Duchowski. Gaze-contingent level of detail rendering. *EuroGraphics (EG)*, 2001. Short Presentation.

[MDK08] Rafał Mantiuk, Scott Daly, and Louis Kerofsky. Display adaptive tone mapping. In *ACM Transactions on Graphics (TOG)*, volume 27, page 68, 2008.

[MDMS05] Rafal Mantiuk, Scott J. Daly, Karol Myszkowski, and Hans-Peter Seidel. Predicting visible differences in high dynamic range images: model and its calibration. In *Electronic Imaging 2005*, pages 204–214. International Society for Optics and Photonics, 2005.

[MET14] Metaio SDK, 2014. http://dev.metaio.com/sdk, vis. 1-5-2015.

[MFN16] Michael Mauderer, David R Flatla, and Miguel A Nacenta. Gaze-contingent manipulation of color perception. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5191–5202. ACM, 2016.

[MG09] Melchi M. Michel and Wilson S. Geisler. 61.1: Invited paper: Gaze contingent displays: Analysis of saccadic plasticity in visual search. In *SID Symposium Digest of Technical Papers*, volume 40, pages 911–914, 2009.

[MHL13] Emilie Møllenbach, John Paulin Hansen, and Martin Lillholm. Eye movements in gaze interaction. *Journal of Eye Movement Research*, 6(2), 2013.

[MHO08] Hiromu Miyashita, Masaki Hayashi, and Ken-ichi Okada. Implementation of eog-based gaze estimation in hmd with head-tracker. In *18th Int. Conf. on Artificial Reality and Telexistence*, pages 20–27, 2008.

[MHP12] Diako Mardanbegi, Dan Witzner Hansen, and Thomas Pederson. Eye-based head gestures. In *Proceedings of the ACM Symposium on Eye-Tracking Research and Applications*, pages 139–146, 2012.

[Mit04] G. E. Mitchell. Taking control over depth of field: Using the lens blur filter in adobe photoshop cs, 2004. http://www.outbackphoto.com/workflow/wf_51/essay.html, vis. 09-12-2016.

[MK06] D. S. Messing and L. J. Kerofsky. Using optimal rendering to visually mask defective subpixels. In *Human Vision and Electronic Imaging XI*, pages 236–247. SPIE, 2006.

[MKRH11] Rafal Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. In *ACM Transactions on Graphics (TOG)*, volume 30, page 40, 2011.

[MM13] Radosław Mantiuk and Mateusz Markowski. Gaze-dependent tone mapping. In *Image Analysis and Recognition*, pages 426–433. Springer, 2013.

[MN84] Suzanne P. McKee and Ken Nakayama. The detection of motion in the peripheral visual field. *Vision research*, 24(1):25–32, 1984.

[MN88] D. P. Mitchell and A. N. Netravali. Reconstruction Filters in Computer-Graphics. In *Proceedings of ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 221–228, 1988.

[Moi11] F. Moisy. Ezyfit: A free curve fitting toolbox for Matlab, 2011. http://www.fast.u-psud.fr/ezyfit/, vis. 09-13-2016.

[MR98] Arien Mack and Irvin Rock. Inattentional blindness: Perception without attention. *Visual attention*, 8:55–76, 1998.

[MR03] Dean R. Melmoth and Jyrki M. Rovamo. Scaling of letter size and contrast equalises perception across eccentricities and set sizes. *Vision Research*, 43(7):769–777, 2003.

[MR06] Norman Murray and Dave Roberts. Comparison of head gaze and head and eye gaze within an immersive environment. In *2006 Tenth IEEE International Symposium on Distributed Simulation and Real-Time Applications*, pages 70–76. IEEE, 2006.

[MRD12] L. McIntosh, Bernhard E. Riecke, and Steve DiPaola. Efficiently simulating the bokeh of polygonal apertures in a post-process depth of field shader. In *Computer Graphics Forum*, volume 31, pages 1810–1822, 2012.

[MRW96] Sabira K. Mannan, Keith H. Ruddock, and David S. Wooding. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial vision*, 10(3):165–188, 1996.

[MS92] Michael E. McCauley and Thomas J. Sharkey. Cybersickness: Perception of self-motion in virtual environments. *Presence: Teleoperators & Virtual Environments*, 1(3):311–318, 1992.

[MTT04] Julio C Martinez-Trujillo and Stefan Treue. Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, 14(9):744–751, 2004.

[Mul85] Kathy T. Mullen. The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of Physiology*, 359(1):381–400, 1985.

[MVOP11] Naila Murray, Maria Vanrell, Xavier Otazu, and C Alejandro Parraga. Saliency estimation using a non-parametric low-level vision model. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 433–440. IEEE, 2011.

[MVS+08] Fiona Mulvey, Arantxa Villanueva, David Sliney, Robert Lange, Sarah Cotmore, and Mick Donegan. Exploration of safety issues in eyetracking. Technical report, Technische Universität Dresden, 2008.

[MWDG13] Belen Masia, Gordon Wetzstein, Piotr Didyk, and Diego Gutierrez. A survey on computational displays: Pushing the boundaries of optics, computation, and perception. *Computers & Graphics*, 37(8):1012–1038, 2013.

[Mys98] Karol Myszkowski. The visible differences predictor: Applications to global illumination problems. In *Rendering Techniques' 98*, pages 223–236. Springer, 1998.

[NAB+15] Rahul Narain, Rachel A. Albert, Abdullah Bulbul, Gregory J. Ward, Martin S. Banks, and James F. O'Brien. Optimal presentation of imagery with focus cues on multi-plane displays. *ACM Transactions on Graphics (TOG)*, 34(4):59, 2015.

[NCR15] Nvidia Corporation and Nathan Reed. Gameworks VR. *Technical Slides*, 2015. https://developer.nvidia.com/sites/default/files/akamai/gameworks/vr/GameWorks_VR_2015_Final_handouts.pdf, vis. 09-12-2016.

[NCRP16] Debanga R. Neog, João L. Cardoso, Anurag Ranjan, and Dinesh K. Pai. Interactive gaze driven animation of the eye region. In *Proceedings of the 21st International Conference on Web3D Technology*, pages 51–59. ACM, 2016.

[NE15] Antje Nuthmann and Wolfgang Einhäuser. A new approach to modeling the influence of image features on fixation selection in scenes. *Annals of the New York Academy of Sciences*, 1339(1):82–96, 2015.

[New16] MIT News. Control your smartphone with your eyes, June 2016. https://www.technologyreview.com/s/601789/control-your-smartphone-with-your-eyes/, vis. 09-26-2016.

[NH12] Antje Nuthmann and John M. Henderson. Using crisp to model global characteristics of fixation durations in scene viewing and reading with a common mechanism. *Visual Cognition*, 20(4-5):457–494, 2012.

[NI02] Vidhya Navalpakkam and Laurent Itti. A goal oriented attention guidance model. In *International Workshop on Biologically Motivated Computer Vision*, pages 453–461. Springer, 2002.

[NI07] Vidhya Navalpakkam and Laurent Itti. Search goal tunes visual features optimally. *Neuron*, 53(4):605–617, 2007.

[NNB+04] Stavri G. Nikolov, Timothy D. Newman, Dave R. Bull, Nishan C. Canagarajah, Michael G. Jones, and Iain D. Gilchrist. Gaze-contingent display using texture mapping and opengl: system and applications. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, pages 11–18. ACM, 2004.

[NSEH10] Antje Nuthmann, Tim J. Smith, Ralf Engbert, and John M. Henderson. Crisp: a computational model of fixation durations in scene viewing. *Psychological Review*, 117(2):382, 2010.

[NSG11] Fernando Navarro, Francisco J. Serón, and Diego Gutierrez. Motion blur rendering: State of the art. *Comput. Graph. Forum*, 30(1):3–26, 2011.

[NSMB+12] AKD Nguyen, AA Simard-Meilleur, C Berthiaume, R Godbout, and L Mottron. Head circumference in canadian male adults: development of a normalized chart. *Int J Morphol*, 30:1474–1480, 2012.

[OCU15] Oculus VR Oculus Rift, 2015. http://oculus.com, vis. 03-12-2015.

[OCV15] OpenCV Library, 2015. http://opencv.org, vis. 03-12-2015.

[OHM+04] Carol O'Sullivan, Sarah Howlett, Yann Morvan, Rachel McDonnell, and Keith O'Conor. Perceptually adaptive graphics. *Eurographics state of the art reports*, 4:1–24, 2004.

[OS95] M.F. O'Brien and N. Sibley. *The Photographic Eye SE: Learning to See with a Camera*. Studio Textbooks Series. Davis Publications, Incorporated, 1995.

[Ost35] G. Osterberg. Topography of the layer of rods and cones in the human retina. *Acta Ophthal. Suppl.*, 1(6):11–97, 1935.

[OTCH03] Aude Oliva, Antonio Torralba, Monica S. Castelhano, and John M. Henderson. Top-down control of visual attention in object detection. In *Image processing, 2003. icip 2003. proceedings. 2003 international conference on*, volume 1, pages I–253. IEEE, 2003.

[Ove10] M.L. Overton. HANSO: Hybrid Algorithm for Non-Smooth Optimization 2.0, 2010. http://www.cs.nyu.edu/overton/software/hanso/, vis. 09-12-2016.

[OYT96] Toshikazu Ohshima, Hiroyuki Yamamoto, and Hideyulu Tamura. Gaze-directed adaptive rendering for interacting with virtual space. In *Virtual Reality Annual International Symposium, 1996., Proceedings of the IEEE 1996*, pages 103–110. IEEE, 1996.

[PCC92] Randy Pausch, Thomas Crea, and Matthew Conway. A literature survey for virtual environments: Military flight simulator visual systems and simulator sickness. *Presence: Teleoperators & Virtual Environments*, 1(3):344–363, 1992.

[PDBB13] Laura Pomarjanschi, Michael Dorr, Peter J. Bex, and Erhardt Barth. *Simple gaze-contingent cues guide eye movements in a realistic driving simulator*, volume 8651. 2013.

[PFD05] Hao Pan, Xiao-Fan Feng, and Scott J. Daly. Lcd motion blur modeling and analysis. In *Proc. of International Conference on Image Processing*, pages 21–24, 2005.

[PG02] Jeffrey S. Perry and Wilson S. Geisler. Gaze-contingent real-time simulation of arbitrary visual fields. In *Electronic Imaging 2002*, pages 57–69. International Society for Optics and Photonics, 2002.

[PKC15] Jakub Pietrzak, Krzysztof Kacperski, and Marek Cieślar. Nvidia optix ray-tracing engine as a new tool for modelling medical imaging systems. In *SPIE Medical Imaging*, pages 94122P–94122P. International Society for Optics and Photonics, 2015.

[Pla00] J.C. Platt. Optimal filtering for patterned displays. *Signal Processing Letters*, pages 179–181, 2000.

[PLN00] Derrick Parkhurst, Irwin Law, and Ernst Niebur. Evaluating gaze-contingent level of detail rendering of virtual environments using visual search. In *Symposium on Eye Tracking Research and Applications (ETRA)*, pages 105–109, November 2000.

[PN03] Derrick J. Parkhurst and Ernst Niebur. Scene content selected by active vision. *Spatial vision*, 16(2):125–154, 2003.

[PN04] Derrick Parkhurst and Ernst Niebur. A feasibility test for perceptually adaptive level of detail rendering on desktop systems. In *Proceedings of the 1st ACM Symposium on Applied Perception (SAP)*, pages 49–56, 2004.

[Por02] Thomas Conrad Porter. Contributions to the study of flicker. paper ii. *Proceedings of the Royal Society of London*, 70(459-466):313–329, 1902.

[Por15] Statista The Statistics Portal, 2015. http://www.statista.com/statistics/485280/flat-panel-tv-area-demand-by-resolution/, vis. 09-16-2015.

[PP99] Jan Prikryl and Werner Purgathofer. Perceptually-driven termination for stochastic radiosity. In *Seventh International Conference in Central Europe on Computer Graphics and Visualization (Winter School on Computer Graphics)*, 1999.

[Pro08] N.T. Proferes. *Film Directing Fundamentals: See Your Film Before Shooting*. Focal Press, 2008.

[PSL99] Dale Purves, Amita Shimpi, and Beau R. Lotto. An empirical explanation of the cornsweet effect. *Journal of Neuroscience*, pages 8542–8551, 1999.

[PT08] Marco Porta and Matteo Turina. Eye-s: a full-screen input modality for pure eye-based communication. In *Proceedings of the 2008 symposium on Eye Tracking Research & Applications*, pages 27–34. ACM, 2008.

[PTYG00] Sumanta N. Pattanaik, Jack Tumblin, Hector Yee, and Donald P. Greenberg. Time-dependent visual adaptation for fast realistic image display. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 47–54. ACM, 2000.

[PVT+13] Ken Pfeuffer, Melodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. Pursuit calibration: Making gaze calibration less tedious and more flexible. In *UIST'13, Proceedings*, pages 261–270. ACM, 2013.

[PY02] Sumanta Pattanaik and Hector Yee. Adaptive gain control for high dynamic range image display. In *Proceedings of the 18th ACM Spring Conference on Computer Graphics*, pages 83–87, 2002.

[Ray92] Keith Rayner. *Eye movements and visual cognition: Scene perception and reading*. Springer Science & Business Media, 1992.

[RB79] Keith Rayner and James H Bertera. Reading without a fovea. *Science*, 206(4417):468–469, 1979.

[RDVŽ10] Snježana Rimac-Drlje, Mario Vranješ, and Drago Žagar. Foveated mean squared error—a novel video quality metric. *Multimedia tools and applications*, 49(3):425–445, 2010.

[Red01] Martin Reddy. Perceptually optimized 3d graphics. *IEEE computer Graphics and Applications*, (5):68–75, 2001.

[Res16] SR Research. SR Research EyeLink 1000, 2016. http://www.sr-research.com/, vis. 09-12-2016.

[Rey98] W. Rey, J.J. On generating random numbers, with help of y = [(a + x)sin(bx)] mod. *Vilnius Conference on Probability Theory and Mathematical Statistics*, 22(7), 1998.

[RFWB07] Ganesh Ramanarayanan, James Ferwerda, Bruce Walter, and Kavita Bala. Visual equivalence: towards a new standard for image fidelity. *ACM Transactions on Graphics (TOG)*, 26(3):76, 2007.

[RG09] Christoph Rasche and Karl R Gegenfurtner. Precision of speed discrimination and smooth pursuit eye movements. *Vision research*, 49(5):514–523, 2009.

[RJG+14] Ryan V. Ringer, Aaron P. Johnson, John G. Gaspar, Mark B. Neider, James Crowell, Arthur F. Kramer, and Lester C. Loschky. Creating a new dynamic measure of the useful field of view using gaze-contingent displays. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 59–66. ACM, 2014.

[RKA16] Raja Koduri and AMD. AMD's Raja Koduri says that we need 16k at 240hz for "true immersion" in VR. Tech-Blog OC3D.NET, 2016. https://t.co/0ae5Mb53WZ, vis. 02-26-2016.

[RKLC+11] Jonathan Ragan-Kelley, Jaakko Lehtinen, Jiawen Chen, Michael Doggett, and Frédo Durand. Decoupled sampling for graphics pipelines. *ACM Transactions on Graphics (TOG)*, 30(3):17, 2011.

[RLMS03] Eyal M. Reingold, Lester C. Loschky, George W. McConkie, and David M. Stampe. Gaze-contingent multiresolutional displays: An integrative review. *HFES Journal*, 45(2):307–328, 2003.

[RMK07] Leila Reddy, Farshad Moradi, and Christof Koch. Top–down biases win against focal attention in the fusiform face area. *Neuroimage*, 38(4):730–739, 2007.

[RR88] Jyrki Rovamo and Antti Raninen. Critical flicker frequency as a function of stimulus area and luminance at various eccentricities in human cone vision: a revision of granit-harper and ferry-porter laws. *Vision research*, 28(7):785–790, 1988.

[RSSF02] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 267–276, 2002.

[RV79] J. Rovamo and V. Virsu. An estimation and application of the human cortical magnification factor. *Experimental brain research*, 37(3):495–510, 1979.

[RVN78] Jyrki Rovamo, Veijo Virsu, and Risto Näsänen. Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature*, 271:54–56, 1978.

[RWPD10] Erik Reinhard, Greg Ward, Sumanta Pattanaik, and Paul Debevec. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2010.

[RYS+16] Jihoon Ryoo, Kiwon Yun, Dimitris Samaras, Samir R. Das, and Gregory Zelinsky. Design and evaluation of a foveated video streaming service for commodity client devices. In *Proceedings of the 7th International Conference on Multimedia Systems*, page 6. ACM, 2016.

[SC02] William R. Sherman and Alan B. Craig. *Understanding virtual reality: Interface, application, and design*. Elsevier, 2002.

[SC06] Veronica Sundstedt and Alan Chalmers. Evaluation of perceptually-based selective rendering techniques using eye-movements analysis. In *Proceedings of the 22nd Spring Conference on Computer Graphics*, pages 153–160. ACM, 2006.

[Sch94] Christophe Schlick. An inexpensive BRDF model for physically-based rendering. In *Computer Graphics Forum*, volume 13, pages 233–246, 1994.

[Sch14] Andrew Schall. *Eye tracking in user experience design*. Morgan Kaufmann, 2014.

[SCM15] Nicholas T. Swafford, Darren Cosker, and Kenny Mitchell. Latency aware Foveated Rendering in Unreal Engine 4. In *Proceedings of the 12th European Conference on Visual Media Production*, page 17. ACM, 2015.

[SD12] Sophie Stellmach and Raimund Dachselt. Look & touch: gaze-supported target acquisition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2981–2990. ACM, 2012.

[SDL+05] Veronica Sundstedt, Kurt Debattista, Peter Longhurst, Alan Chalmers, and Tom Troscianko. Visual attention for efficient high-fidelity graphics. In *Proceedings of the 21st Spring Conference on Computer Graphics*, pages 169–175. ACM, 2005.

[Sha49] Claude Elwood Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[SHH07] Frank Scharnowski, Frouke Hermens, and Michael H Herzog. Bloch's law and the dynamics of feature fusion. *Vision research*, 47(18):2444–2452, 2007.

[SI10] John Shen and Laurent Itti. Gender differences in visual attention during listening as measured by neuromorphic saliency: What women (and men) watch. *Journal of Vision*, 10(7):159–159, 2010.

[SIGK+16] Nicholas T. Swafford, José A. Iglesias-Guitian, Charalampos Koniaris, Bochang Moon, Darren Cosker, and Kenny Mitchell. User, metric, and computational evaluation of foveated rendering methods. *ACM Symposium on Applied Perception (SAP)*, July 2016.

[SLL+14] Christian Scheel, Falko Löffler, Anke Lehmann, Heidrun Schumann, and Oliver Staadt. Dynamic level of detail for tiled large high-resolution displays. *GI VR/AR Fachtagung*, 2014.

[SM16] Michael Stengel and Marcus Magnor. Gaze-contingent computational displays. *IEEE Signal Processing Magazine (SPM)*, 33(5):139–148, September 2016.

[SMI16] SMI Eye Tracking HMD Upgrade for the Oculus Rift DK2, 2016. http://www.smivision.com/, vis. 03-28-2016.

[SNRS12] Daniel Scherzer, Chuong H. Nguyen, Tobias Ritschel, and Hans-Peter Seidel. Pre-convolved radiance caching. In *Computer Graphics Forum*, volume 31, pages 1391–1397, 2012.

[Sol16] Solid Angle. Arnold renderer, 2016. http://solidangle.com, vis. 09-12-2016.

[SRIR07] Fabrizio Santini, Gabriel Redner, Ramon Iovin, and Michele Rucci. Eyeris: a general-purpose system for eye-movement-contingent display control. *Behavior Research Methods*, 39(3):350–364, 2007.

[SRJ11] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):13, 2011.

[SSTT12] Robert Snowden, Robert J Snowden, Peter Thompson, and Tom Troscianko. *Basic vision: an introduction to visual perception*. Oxford University Press, 2012.

[ST90] Takafumi Saito and Tokiichiro Takahashi. Comprehensible rendering of 3-d shapes. In *ACM SIGGRAPH Computer Graphics*, volume 24, pages 197–206. ACM, 1990.

[STNE15] Josef Stoll, Michael Thrun, Antje Nuthmann, and Wolfgang Einhäuser. Overt attention in natural scenes: objects dominate features. *Vision research*, 107:36–48, 2015.

[SU07] Jonathan A. Stirk and Geoffrey Underwood. Low-level visual saliency does not predict change detection in natural scenes. *Journal of vision*, 7(10):3–3, 2007.

[SVV+09] Erich Schneider, Thomas Villgrattner, Johannes Vockeroth, Klaus Bartl, Stefan Kohlbecher, Stanislavs Bardins, Heinz Ulbrich, and Thomas Brandt. Eyeseecam: An eye movement–driven head camera for the examination of natural visual exploration. *Annals of the NY Academy of Science*, 1164(1):461–467, 2009.

[SW14] Daniel R. Saunders and Russell L. Woods. Direct measurement of the system latency of gaze-contingent displays. *Behavior research methods*, 46(2):439–447, 2014.

[SWM+08] William Steptoe, Robin Wolff, Alessio Murgia, Estefania Guimaraes, John Rae, Paul Sharkey, David Roberts, and Anthony Steed. Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments. In *Proceedings of the 2008 ACM Conference on Computer-supported Cooperative Work*, pages 197–200. ACM, 2008.

[SYGM03] J. Stewart, J. Yu, S. J. Gortler, and L. McMillan. A new reconstruction filter for undersampled light fields. In *Proceedings of the 14th Eurographics Workshop on Rendering*, pages 150–156, 2003.

[SYM+11] Daniel Scherzer, Lei Yang, Oliver Mattausch, Diego Nehab, Pedro V. Sander, Michael Wimmer, and Elmar Eisemann. A survey on temporal coherence methods in real-time rendering. In *EUROGRAPHICS 2011 State of the Art Reports*, pages 101–126. Eurographics Association, 2011.

[TA13] Cihan Topala and Cuneyt Akinlara. An adaptive algorithm for precise pupil boundary detection using the entropy of contour gradients. *Elsevier*, 2013. Elsevier preprint.

[TDM+14] Krzysztof Templin, Piotr Didyk, Karol Myszkowski, Mohamed M. Hefeeda, Hans-Peter Seidel, and Wojciech Matusik. Modeling and optimizing eye vergence response to stereoscopic cuts. *ACM Transactions on Graphics (TOG)*, 33(4):145:1–145:8, July 2014.

[TDMS16] Krzysztof Templin, Piotr Didyk, Karol Myszkowski, and Hans-Peter Seidel. Emulating displays with continuously varying frame rates. *ACM Transactions on Graphics (TOG)*, 35(4):67, 2016.

[TDR+11] Krzysztof Templin, Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, and Hans-Peter Seidel. Apparent resolution enhancement for animations. In *Proceedings of ACM Spring Conference on Computer Graphics (SCCG)*, pages 85–92, 2011.

[TFCRS11] William Thompson, Roland Fleming, Sarah Creem-Regehr, and Jeanine Kelly Stefanucci. *Visual perception from a computer graphics perspective*. CRC Press, 2011.

[Thi89] Larry N. Thibos. Image processing by the human eye. In *1989 Advances in Intelligent Robotics Systems Conference*, pages 1148–1153. International Society for Optics and Photonics, 1989.

[Tin14] Angela Tinwell. *The Uncanny Valley in Games and Animation*. CRC Press, 2014.

[TKA02] Kar-Han Tan, David Kriegman, and Narendra Ahuja. Appearance-based eye gaze estimation. In *WACV 2002, Proceedings*, pages 191–195, 2002.

[TL14] Wen-Jiin Tsai and Yi-Shih Liu. Foveation-based image quality assessment. In *Visual Communications and Image Processing Conference, 2014 IEEE*, pages 25–28. IEEE, 2014.

[Tre88] Anne Treisman. Features and objects: The fourteenth bartlett memorial lecture. *The quarterly journal of experimental psychology*, 40(2):201–237, 1988.

[Tru13] Douglas Trumbull. Douglas Trumbull On High Frame Rate Filmmaking, 2013. http://www.moviescopemag.com/market-news/featured-editorial/douglas-trumbull-on-high-frame-rate-filmmaking-part-1/, 08-23-2014.

[TSY+07] Jacob Telleen, Anne Sullivan, Jerry Yee, Oliver Wang, Prabath Gunawardane, Ian Collins, and James Davis. Synthetic shutter speed imaging. *Comput. Graph. Forum*, 26(3):591–598, 2007.

[TW01] Lisa C. Thomas and Christopher D. Wickens. Visual displays and cognitive tunneling: Frames of reference effects on spatial judgments and change detection. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 45, pages 336–340. SAGE Publications, 2001.

[TZS+16] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1, 2016.

[Und98] Geoffrey Underwood. *Eye guidance in reading and scene perception*. Elsevier, 1998.

[Upg14] SMI Oculus Rift Upgrade, November 2014. http://newatlas.com/eye-tracking-oculus-rift/34878/, vis. 09-19-2016.

[VAF16] M. Vinnikov, R. S. Allison, and S. Fernandes. Impact of depth of field simulation on visual fatigue: Who are impacted? and how? *International Journal of Human-Computer Studies*, 91:37–51, 2016.

[VAS08] Margarita Vinnikov, Robert S. Allison, and Dominik Swierad. Real-time simulation of visual defects with gaze-contingent display. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, pages 127–130. ACM, 2008.

[VCD09] Suzane Vassallo, Sian L Cooper, and Jacinta M Douglas. Visual scanning in the recognition of facial affect: Is there an observer sex difference? *Journal of Vision*, 9(3):11–11, 2009.

[VDC14] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805, 2014.

[VDMB12] Eleonora Vig, Michael Dorr, Thomas Martinetz, and Erhardt Barth. Intrinsic dimensionality predicts the saliency of natural dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1080–1091, 2012.

[VG07] Roger P. G. Van Gompel. *Eye movements: A window on mind and brain*. Elsevier, 2007.

[vH05] Hans van Hateren. A cellular and molecular model of response kinetics and adaptation in primate cones and horizontal cells. *Journal of Vision*, pages 331–347, 2005.

[VJ04] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, pages 137–154, 2004.

[Vla16] Alex Vlachos. Advanced VR Rendering Performance. Game Developer Confernce (GDC) 2016, March 2016. http://alex.vlachos.com/graphics/Alex_Vlachos_Advanced_VR_Rendering_Performance_GDC2016.pdf, vis. 09-12-2016.

[VNO87] Veijo Virsu, Risto Näsänen, and Kari Osmoviita. Cortical magnification and peripheral vision. *JOSA A*, 4(8):1568–1578, 1987.

[VST+14] Karthik Vaidyanathan, Marco Salvi, Robert Toth, Tim Foley, Tomas Akenine-Möller, Jim Nilsson, Jacob Munkberg, Jon Hasselgren, Masamichi Sugihara, Petrik Clarberg, et al. Coarse pixel shading. In *Eurographics/ACM SIGGRAPH Symposium on High Performance Graphics*, pages 9–18. The Eurographics Association, 2014.

[VWSC03] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In *Proceedings of the ACM SIGCHI conference on Human factors in computing systems*, pages 521–528, 2003.

[Wan95] B. A. Wandell. Useful quantities in vision science. In *Foundations of Vision*. Sinauer Associates Inc Sunderland, Massachusetts, 1995.

[Wat13] Andrew B. Watson. High frame rates and human vision: A view through the window of visibility. *SMPTE Motion Imaging Journal*, 122(2):18–32, 2013.

[WBC06] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the S3GP. In *CVPR'06, Proceedings*, volume 1, pages 230–237, 2006.

[WBLK01] Zhou Wang, Alan C. Bovik, Ligang Lu, and Jack L. Kouloheris. Foveated wavelet image quality index. In *International Symposium on Optical Science and Technology*, pages 42–52. International Society for Optics and Photonics, 2001.

[WBLP11] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Real-time performance-based facial animation. In *ACM Transactions on Graphics (TOG)*, volume 30, page 77, 2011.

[WBSS04] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[WDW99] A. Mark Williams, Keith Davids, and John Garrett Pascoe Williams. *Visual perception and action in sport*. Taylor & Francis, 1999.

[WDZW15] Xuyang Wang, Yangdong Deng, Guiju Zhang, and Zhihua Wang. Apparent resolution enhancement for near-eye light field display. In *SIGGRAPH Asia 2015 Mobile Graphics and Interactive Applications*, page 4. ACM, 2015.

[Wey63] Frank W. Weymouth. Visual sensory units and the minimum angle of resolution. *Optometry & Vision Science*, 40(9):550–568, 1963.

[WHLP16] Veronica U. Weser, Joel Hesch, Johnny Lee, and Dennis R. Proffitt. User sensitivity to speed-and height-mismatch in VR. In *Proceedings of the ACM Symposium on Applied Perception (SAP)*, pages 143–143, 2016.

[Wil85] David R. Williams. Aliasing in human foveal vision. *Vision Research*, 25(2):195–205, 1985.

[WK06] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural networks*, 19(9):1395–1407, 2006.

[WLC+03] Nathaniel Williams, David Luebke, Jonathan D. Cohen, Michael Kelley, and Brenden Schubert. Perceptually guided simplification of lit, textured meshes. In *Proceedings of the ACM Symposium on Interactive 3D Graphics*, pages 113–121, 2003.

[WM78] Gerald Westheimer and Suzanne P. McKee. Stereoscopic acuity for moving retinal images. *JOSA*, 68(4):450–455, 1978.

[WMPH16] Dong Wang, Fiona B. Mulvey, Jeff B. Pelz, and Kenneth Holmqvist. A study of artificial eyes for the measurement of precision in eye-trackers. *Behavior Research Methods*, pages 1–13, 2016.

[WRSD08] Jacob O. Wobbrock, James Rubinstein, Michael W. Sawyer, and Andrew T. Duchowski. Longitudinal evaluation of discrete consecutive gaze gestures for text entry. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, pages 11–18. ACM, 2008.

[WTP+09] Manuel Werlberger, Werner Trobin, Thomas Pock, Andreas Wedel, Daniel Cremers, and Horst Bischof. Anisotropic Huber-L^1 optical flow. In *Proc. of the British Machine Vision Conference (BMVC)*, pages 1–11, 2009.

[YIMS08] Akiko Yoshida, Matthias Ihrke, RafałMantiuk, and Hans-Peter Seidel. Brightness of the glare illusion. In *Proceedings of the ACM Symposium on Applied Perception in Graphics and Visualization (APGV)*, pages 83–90, 2008.

[YPG01] Hector Yee, Sumanita Pattanaik, and Donald P. Greenberg. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Transactions on Graphics (TOG)*, 20(1):39–65, 2001.

[ZK13] Qi Zhao and Christof Koch. Learning saliency-based visual attention: A review. *Signal Processing*, 93(6):1401–1407, 2013.

[ZPB07] C. Zach, T. Pock, and H. Bischof. A duality based approach for real-time TV-L1 optical flow. In *Proceedings of the 29th DAGM Conference on Pattern Recognition*, pages 214–223, 2007.

[ZR12] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886. IEEE, 2012.

[ZS06] Yun Zhai and Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, pages 815–824, 2006.

Glossary

ACCOMMODATION The ability of changing the focal distance of the eye's lens. This is done to bring a new object into focus, normally in coordination with vergence eye movements (convergence and divergence).

ADAPTATION Adaptation is the automatically triggered and time-dependent process of tuning sensitivity of retinal cells and neurons to the amount of incoming light.

ALIASING Aliasing may result in visual artifacts caused by undersampling a signal spatially or temporally. Spatial aliasing may result in jagged edges whereas temporal aliasing results in animation artifacts, such as ghosting, incomplete plane rotors or wheels apparently rotating backwards.

AMOLED See *light-emitting diode*.

ANTI-ALIASING Strategies for removing or reducing *aliasing artifacts* arising from spatial or temporal undersampling.

AR See *augmented reality*.

AUGMENTED REALITY (AR) In an AR experience the user perceives the real and virtual world combined and at the same time. Virtual objects appear fixed in space and can be interacted with (Azuma 1997).

AVATAR The representation of a user in a virtual environment. Depending on the intention and complexity of the application the representation may be physically plausible or intentionally abstract.

BINOCULAR Using two eyes.

BINOCULAR DEPTH CUES See *depth cues*.

BINOCULAR DISPARITY The differences in perspective between the viewpoints of the left and right eye. The binocular disparity has to match the IPD for realistic depth impression of the virtual environment (see *inter-pupillary distance*).

BINOCULAR VISION Viewing with two eyes. Both views are fused by the brain resulting a three-dimensional scene representation.

BOUNDING VOLUME A bounding volume that completely contains an object; typically a box or a sphere.

CFF See *critical flicker frequency*.

COLLECTOR CELLS The group of cells in the retina that lie between the *photoreceptor cell*s and *retinal ganglion cells*.

CONE A color-sensitive photoreceptor in the human retina (see *photopic vision*).

CONTRAST The light intensity difference received at one point of the retina and its local surrounding.

CONTRAST GRATING An alternating pattern of bright and dark bars. Used to measure a subject's contrast sensitivity.

CONTRAST SENSITIVITY The reciprocal of threshold contrast measuring the subject's sensitivity to spatial detail. Measurements over spatial frequency and contrast result in the Contrast Sensitivity Function (CSF).

CONVERGENCE Ability of the eyes to move inwards. See *accommodation*.

CORNEA The transparent matter at the frontal part of the eye.

CORTICAL MAGNIFICATION FACTOR (CMF) Describes the drop-off in retinal sensitivity out toward the peripheral field. This factor is often given the label M. It has been shown that M^2 is directly proportional to the density of receptive fields of retinal ganglion cells.

CRITICAL FLICKER FREQUENCY (CFF) The frame rate at which a sequentially presented series of images appears continuous, or is perceptually fused, synonymously named *critical flicker frequency*. Measured in Hertz (Hz). For most people, the CFF is roughly 70 Hz. The CFF in the visual periphery may be even higher under non-default environment illumination.

CSF See *contrast sensitivity*.

DARK ADAPTATION Describes the adjustment process of the light-adapted eye to a dark environment.

DEFERRED SHADING The term coins a rendering technique which splits visibility computation from shading. This allows to simulate thousands of dynamic lights in a highly complex simulated scene.

DELAY See *latency*.

DEPTH BUFFER The depth buffer (also z-buffer) is a memory buffer holding depth values representing the distance between camera and scene geometry. The buffer enables fast hardware-supported depth tests being essential for most drawing and shading tasks.

DEPTH CUES Strategies such as eye convergence (binocular depth cue), motion parallax (*monocular* depth cue) and perspective for estimating the distance of an object.

DIVERGENCE Ability of the eyes to move outwards. See *accomodation*.

ECCENTRICITY Angular deviation from the center of the fovea.

ETHMD See *eye-tracking head-mountd display*.

EYE TRACKING Capturing the gaze direction of one eye (*monocular* eye tracking) or both eyes (*binocular* eye tracking).

EYE-TRACKING HEAD-MOUNTED DISPLAY (ETHMD) A VR headset (head-mounted display) with integrated gaze estimation functionality.

FIELD OF VIEW (FOV) The solid angular region that is visible to the eye.

FIXATION Gazing at a point of the scene or display for a certain time (fixation duration).

FOV See *field of view*.

FOVEA The retinal area able to perceive a visual information at highest detail.

FRAME A complete image in an animated (mono or stereo) sequence. Cinematic film typically use 24 frames per second. Every element of a frame represents the same moment in time.

FRAME RATE The number of frames displayed in a certain amount of time. Typically measured in frames per second (fps) or Hertz (Hz).

FRAME TIME The duration, or time of display, for one frame. Frame time is the inverse of frame rate. Typically measured in milliseconds (ms).

GAZE DIRECTION The viewer's eye direction.

HEAD TRACKING Measuring the location (position and orientation) of the user's head with respect to a global reference frame.

HEAD-MOUNTED DISPLAY (HMD) A wearable display mounted in front of the user's eyes for *VR* or *AR* applications.

HFR See *high frame rate video*.

HIGH FRAME RATE VIDEO (HFR) Video frame rates of 48 – 60 Hz; successor of traditional frame rate enables reduced motion blur and smoother motion perception in movies and immersive experiences.

HIGH-LEVEL PERCEPTION The "top-down" processing of the *human visual system*. High-level perception is concerned with how known objects are recognized. See also *low-level perception*.

HMD See *head-mounted display*.

HUMAN VISUAL SYSTEM (HVS) The HVS abstracts those parts of the human body which are responsible for visual information processing. The common HVS model comprises the eyes, the connecting visual pathways and the visual cortex of the brain.

HVS See *human visual system*.

HYPERACUITY Perception of features that are smaller than the spacing of a photoreceptor cells.

IMAGE PYRAMID A data structure used for efficient filtering of images and texture data.

IMMERSION A term for the sensation of being in an environment. This can be a purely mental state (mental immersion) or can be accomplished through physical means (sensory immersion).

IMU See *inertial measurement unit*.

INDEX OF REFRACTION (IOR) The index of refraction (also refractive index) is a material-specific number describing how much light is bent or refracted when light enters that material.

INERTIAL MEASUREMENT UNIT (IMU) A multi-sensor device used for measuring the orientation of objects in relation to a calibrated world frame. Typically, an IMU integrates data from an accelerometer, a gyroscope and a magnetometer.

INFRARED LIGHT (IR) The part of the light spectrum from 700 nm to 1mm being invisible to human visual perception. Used for unobtrusive active gaze tracking.

INTER-PUPILLARY DISTANCE (IPD) The distance between the optical centers of our eyes. The term *inter-ocular distance* is used synonymously.

IOR See *index of refraction*.

IPD See *inter-pupillary distance*.

IR See *infrared light*.

LAG See *latency*.

LATENCY A term describing the duration from starting computations of an image until photons from the displayed frame hit the user's retina. The terms *delay* and *lag* are used synonymously.

LED See *light-emitting diode*.

LEVEL OF DETAIL (LOD) Rendering objects at different resolutions allows adjusting between rendering performance and image quality with respect to the resulting visual detail in the projected image.

LIGHT ADAPTATION Describes the adjustment process of the dark-adapted eye to a bright environment.

LIGHT-EMITTING DIODE (LED) A light-emitting diode (LED) is a semiconductor light source which emits light of a specific wavelength. LEDs are generally used for lighting, illumination units in displays or as sensors. Organic LEDs (OLED) contain an organic compound and enable displays without a backlight resulting in higher contrast and brighter displays.

LOD See *level of detail*.

LOW FRAME RATE VIDEO (LFR) Video frame rates of 24 – 30 Hz; traditionally used in cinemas and for TV broadcasting.

LOW-LEVEL PERCEPTION The "bottom-level" processing in the early stages of the human visual system. Models allow *saliency* estimation for gaze-contingent rendering and computer vision applications. See also *high-level perception*.

MIP-MAPPING A fast linear texture filtering approach using an image pyramid performed on GPU hardware.

MONOCULAR Using one eye.

MONOCULAR DEPTH CUES See *depth cues*.

MOTION SICKNESS Over time conflicting visual and motion cues result in motion sickness (also known as "nausea" or "simulation sickness").

OBJECT CONSTANCY The way objects in the real world appear stationary although the eyes or the head are in motion.

OBJECT OF INTEREST (OOI) An object or part of a scene the user is looking at. The OOI can be estimated either by using active *eye tracking* or approximated by *saliency analysis*.

OLED See *light-emitting diode*.

OOI See *object of interest*.

PERIPHERAL VISION Visual information detected at the periphery of our field of view.

PHOTOPIC VISION Vision with the use of cone receptors.

PHOTORECEPTORS Include those retinal cells (rods and cones) which convert light received at the retina into nerve signals. Rods are achromatic and sensitive to motion. Cones provide color sensitivity.

PRESENCE The sense of presence means being mentally immersed.

PULL-PUSH An efficient algorithm based on a data pyramid for the interpolation of scattered data.

RAY TRACING Computation of the light transport in a scene based on the the geometry of light rays.

REAL-TIME An almost instantaneous visual reaction to any change of the input to the *rendering* system.

REFRESH RATE The rate at which the display screen is refreshed (measured in Hz).

RENDERING The process of computing for a 3D object a 2D representation which can be displayed.

RESOLUTION A measure of a system's ability to capture or display spatial detail (number of pixels, *cf. spatial resolution*) or temporal detail (frequency in Hz).

RETINAL GANGLION CELLS The output neurons containing circular receptive fields in order to encode and transmit information from the eye to the brain.

RODS Light receptors in the retina that are active in dim lighting conditions (scotopic vision).

SACCADE A rapid reflex movement of the eye which is made in order to fixate a target onto the fovea.

SACCADIC SUPPRESSION The effect that the visual system seems to shut down to some degree during *saccades*. That is, even though the point of fixation moves at very high velocities during a saccade, blurred vision is not experienced.

SALIENCY The perceptual importance of an object in a scene. Saliency estimation in Computer Vision commonly computes a saliency value for each image pixel by models of low-level or high-level properties of *human vision*.

SHADING The process of computing the light transport in a scene based on models of optics and materials.

SPATIAL RESOLUTION A measure of the degree of spatial detail that the eye can perceive. This is normally given in units of cycles per degree of visual arc (c/deg).

SPATIOTEMPORAL THRESHOLD SURFACE The surface that describes the sensitivity of an observer to stimuli of varying spatial and temporal characteristics.

STEREOPSIS See *binocular vision*.

SYSTEM LATENCY (LAG) The time duration required for creating and displaying a visual information via the rendering system. Measured in milliseconds (ms).

TELEPRESENCE The experience of *presence* in an environment by means of a communication medium [**?**].

THRESHOLD CONTRAST The minimum contrast required to see a target.

TRACKING Monitoring an object's position and orientation.

UHFR See *ultra-high frame rate video*.

ULTRA-HIGH FRAME RATE VIDEO (UHFR) Video frame rates above 1000 Hz; enables temporal filtering for almost aliasing-free video playback.

UPDATE RATE See *refresh rate*.

VE See *virtual environment*.

VERGENCE EYE MOTION The synchronized movement of both eyes which, along with accommodation, allows to focus at a point with particular depth.

VESTIBULAR SYSTEM Monitors the body's acceleration, equilibrium and relationship with the earth's gravitational field.

VIEW-DEPENDENT LOD A level-of-detail scheme in which surface detail is varied dynamically, retessellating objects on the fly relative to the user's viewpoint, and continuously, allowing a single object to span multiple levels of detail.

VIRTUAL ENVIRONMENT (VE) The rendered 3D scene which can represent real-world objects or abstract data.

VIRTUAL REALITY (VR) A generic term for systems that create a real-time immersive visual/audio/haptic experience. In other words, VR means an alternate world filled with computer-generated images that respond to human movements (Greeenbaum, 1992). VR is realized by an electronic simulation of environments experienced via head-mounted displays and tracking devices enabling the user to interact in realisitc three-dimensional situations (Coates, 1992).

VISUAL ACUITY Measurement for the ability to see visual detail under ideal illumination conditions. Visual acuity is primarily limited by the optics of the eye and by the physiology of the visual system [AKLA11].

VISUAL CORTEX Part of the brain used for processing visual information.

VISUAL CUES Signals or prompts derived from a scene.

VISUAL MASKING The observation that visual pattern affect each other. Hence, if a pattern is present another pattern may get less visible.

VR See *virtual reality*.

List of Figures

Image Credits

THE WHITE ROOM scene (Fig. 5.7) is courtesy of JAY-ARTIST (CC-BY license).

The SINTEL Durian Open Movie project and BIG BUCK BUNNY (Fig. 4.1 and Fig. 4.6) are courtesy of the Blender Foundation (CC-BY license).

The GARDEN video sequence (Fig. 4.6) is courtesy of Stephen Higgins.

The LUPE video sequence (Fig. 4.6) is courtesy of Evin Grant.

The BIKE sequence (Fig. 5.6) is courtesy of the RED Digital Cinema Camera Company.

The models MATINEE, HOUSE, and AVATAR (Fig. 6.10 and Fig. 6.11) and the Unreal Engine are provided by Epic Games, Inc (Free license for academic use).

The SPONZA model (Fig. 7.1, 7.4 and 7.5) is courtesy of Crytek GmbH. Textures are provided by Alexandre Pestana, Frank Meinl, and Morgan McGuire (CC-BY license).

The NEPTUNE model (Fig. 5.2) is courtesy of 3dcadbrowser.com (CC-BY license).

The SKULL model (Fig. 5.4) is courtesy of ColeHarris (CC-Zero license).

The CERBERUS model (Fig. 7.5) is courtesy of Andrew Maximov (CC-BY license).

Thanks to all artists and companies for offering to use your great work.

Publications

Journal Articles

- Michael Stengel, Martin Eisemann, Stephan Wenger, Benjamin Hell, and Marcus Magnor. **Optimizing Apparent Display Resolution Enhancement for Arbitrary Videos**. In *IEEE Transactions on Image Processing (TIP)*, vol. 22, no. 9, pages 3604–3613, September 2013. Patent number 10 2013 105 638.

- Michael Stengel, Pablo Bauszat, Martin Eisemann, Elmar Eisemann, and Marcus Magnor. **Temporal Video Filtering and Exposure Control for Perceptual Motion Blur**. In *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, vol. 21, no. 5, pages 663–671, May 2015.

- Michael Stengel and Marcus Magnor. **Gaze-contingent Computational Displays**. In *IEEE Signal Processing Magazine (SPM)*, vol. 33, no. 5, pages 139–148, September 2016.

International, Refereed Conferences

- Jan Jacobs, Michael Stengel, and Bernd Fröhlich. **A Generalized God-Object Method for Plausible Finger-Based Interactions in Virtual Environments**. In *Proceedings of IEEE Symposium on 3D User Interfaces (3DUI)*, pages 43–51, March 2012.

- Lorenz Rogge, Felix Klose, Michael Stengel, Martin Eisemann, and Marcus Magnor. **Garment Replacement in Monocular Video Sequences**. In *ACM Transactions on Graphics (TOG)*, vol. 34, no. 1, pages 6:1–6:10, November 2014.

- Michael Stengel, Steve Grogorick, Martin Eisemann, Elmar Eisemann, and Marcus Magnor. **An Affordable Solution for Binocular Eye Tracking and Calibration in Head-mounted Displays**. In *Proceedings of ACM Multimedia MM '15*, pages 15–24, October 2015.

- Thomas Löwe, Michael Stengel, Emmy-Charlotte Förster, Steve Grogorick, Marcus Magnor. **Visualization and Analysis of Head Movement and Gaze Data for Immersive Video in Head-mounted Displays**. In *Proceedings of the Workshop on Eye Tracking and Visualization (ETVIS)*, vol. 1, October 2015.

- Michael Stengel, Steve Grogorick, Martin Eisemann, and Marcus Magnor.
  Adaptive Image-Space Sampling for Gaze-Contingent Real-time Rendering.
  In *Computer Graphics Forum (Proceedings of Eurographics Symposium on Rendering EGSR)*, vol. 35, no. 4, pages 129–139, July 2016.

Book Chapters

- Anna Hilsmann, Michael Stengel, Lorenz Rogge.
  Cloth Modeling. In *Digital Representations of the Real World: How to Capture, Model, and Render Visual Reality* by Marcus Magnor, Oliver Grau, Olga Sorkine-Hornung and Christian Theobalt, pages 229–243, May 2015.

- Thomas Löwe, Michael Stengel, Emmy-Charlotte Förster, Steve Grogorick, Marcus Magnor.
  Gaze Visualization for Immersive Video. In *Eye Tracking and Visualization*, Springer Verlag, to appear, 2016.

Refereed Posters

- Michael Stengel, Steve Grogorick, Lorenz Rogge, and Marcus Magnor.
  A Nonobscuring Eye Tracking Solution for Wide Field-of-View Head-mounted Displays, Eurographics 2014, April 2014.

- Michael Stengel, Steve Grogorick, Martin Eisemann, Elmar Eisemann, and Marcus Magnor.
  A Nonobscuring Eye Tracking Solution for Wide Field-of-View Head-mounted Displays, IEEE Virtual Reality (VR) Conference 2015, March 2015.

- Michael Stengel, Steve Grogorick, Elmar Eisemann, Martin Eisemann, and Marcus Magnor.
  An Affordable Solution for Binocular Eye Tracking and Calibration in Head-mounted Displays, ACM Multimedia 2015, October 2015.

Michael Stengel
September 2016

Curriculum Vitæ - Lebenslauf

Curriculum Vitæ

1985	born in Magdeburg, Germany
2005	High School degree, main subjects physics and mathematics
	Dr.-Frank-Gymnasium Staßfurt, Germany
2005 - 2011	Diploma in Computational Visualistics
	Otto-von-Guericke Universität Magdeburg, Germany
2011 - 2016	Ph.D. Student in Computer Science, Prof. M. Magnor
	TU Braunschweig, Germany

Lebenslauf

1985	geboren in Magdeburg, Deutschland
2005	Abitur, Leistungskurse Mathematik und Physik
	Dr.-Frank-Gymnasium Staßfurt, Deutschland
2005 - 2011	Diplom im Studiengang Computervisualistik
	Otto-von-Guericke Universität Magdeburg, Deutschland
2011 - 2016	Promotionsstudent Informatik, Prof. M. Magnor
	TU Braunschweig, Deutschland

Michael Stengel
September 2016